In the Specification

Please amend the specification by changing the word "step" on page 30 line 32 to –means--.
A marked up paragraph showing the correction is just below.

Marked up paragraph from page 30:

The CL-F region and covering markers are for a species and the one or more individuals are members of the species. Means for determining information on the presence or absence of each allele of each bi-allelic marker of the group in chromosomal DNA includes any means of determination. Means for determining information on the presence or absence of each allele of each bi-allelic marker of the group in chromosomal DNA includes means comprising oligonucleotide technology by using a set of oligonucleotides that is complementary to the group as discussed below. Information on the presence or absence of each allele in the chromosomal DNA is obtained using a DNA specimen from each of one or more individuals of the sample or by using one or more DNA pools of DNA specimens from two or more individuals of the sample. Any apparatus that obtains genotype data or sample allele frequency data (similar to the data of the step d) of process #1) by determining the presence or absence of each allele of each bi-allelic marker of the group in the chromosomal DNA of one or more individuals is an example of this version of the invention. Versions of this apparatus also obtain a combination of genotype data and sample allele frequency data similar to the data of the step d) of process #1. The details of ~~step~~ means b) will be clear to those of ordinary skill in the art.

In the Specification (continued)

Please amend the specification, on page 6 line 10 insert the words – - Muller-Mysok & Abel (1997) independently made a similar observation, but they emphasized the weakness of TDT power when the m/p ratio departs from unity and δ is not close to δ_{max}.--

A marked up paragraph showing the correction is just below.

Marked up paragraph from page 6:

published.[11] In this paper a general framework for determining the power of the TDT in many different situations is presented. The analysis of Risch and Merikangas[8] and others is shown by the inventor to be a special case of his general framework. His observations and calculations published in this paper have shown that the TDT has increased power in more common, less optimal situations as well as the less common, optimal situation cited by Muller-Myshok and Abel.[9] As opposed to the observation of Muller-Myhsok and Abel, the inventor's calculations indicate that association tests such as the TDT have increased power in typical situations even when the ratio m/p departs significantly from unity and, or the linkage disequilibrium between the analyzed (marker) allele and disease polymorphism is only half its maximum possible value. The inventor arrived at these conclusions independently and did not derive them from others. Muller-Mysok & Abel (1997) independently made a similar observation, but they emphasized the weakness of TDT power when the m/p ratio departs from unity and δ is not close to δ_{max}.

Remarks

The amendment on page 6 is the incorporation of matter from the last paragraph of page 166 of the inventor's paper. This paper is incorporated by reference into the patent application and is cited in foot note 11 on page 6.

The amendment on page 30 is the correction of an error that a person of ordinary skill in the art would immediately recognize as an error.

Also included herewith is new evidence/arguments regarding patentability. It is given below.

New evidence/arguments

The applicants intend to submit new claims that depend from claims for which the applicants have received a Notice of Allowance (dated 18 April 2003). The applicants intend to submit these new claims before the expiration of the 3 month time period constituting the requested period for limited Suspension of Action under 37CFR 1.103(c).

It is also the wish of the applicants to pursue apparatus claims in future prosecution. Also enclosed are copies of eight publications. These copies of publications are being sent to the USPTO as evidence of the patentability of apparatus claims/embodiments and of the support for such apparatus claims/embodiments in the patent application. More specifically the Examiner stated in the last paragraph of page 4 of the Final Office Action dated 02 OCT 2002 that should the applicants wish to pursue apparatus claims in future prosecution that concrete examples are needed in the specification in order for "means plus function" language to be interpreted and searched. Therefore these eight publications are being provided as evidence in support of patentability.

Each of these publications is cited in the specification of the patent application and is incorporated by reference into the application. More specifically see some of the concrete examples described on p. 24 lines 1 to 2, p. 29 lines 28 to 30 and p. 34 lines 3 to 18, of the patent application. Each of these sections of the application (and the associated publications in the endnotes) describe some concrete examples of technology in the patent application for the interpretation of "means plus function" language. These sections of the application also refer to publications in the endnotes. The publications in the endnotes are incorporated by reference into the patent application.

The copies of eight publications enclosed herein are the publications listed below:

1) Weighing DNA for Fast Genetic Diagnosis, Science, March 27, 1998, vol. 279, pp. 2044-2045.
2) Accessing Genetic Information with High-Density DNA Arrays, Mark Chee, et al. Science, vol 274, Oct. 25, 1996 , pp. 610 – 614.
3) Genetic analysis of amplified DNA with immobilized sequence- specific oligonucleotide probes, Saiki,et al. Proc Natl Acad Sci USA vol 86, pp.6230-6234.
4) Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia, Wu, et al., Proc Natl Acad Sci USA vol 86 pp 2757-2760.
5) Automated DNA diagnostics using an Elisa-based oligonucleotide ligation assay, Nickerson, et al., Proc Natl Acad Sci USA vol 87, pp. 8923-8927.
6) Padlock Probes:Circularizing Oligonucleotides for Localized DNA Detection, Science, Sept. 30, 1994, vol. 265, pp. 2085-2088.
7) SNP attack on complex traits, Nature Genetics, Nov. 1998, vol. 20 no. 3, pp. 217-218.
8) Large Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome, Wang, et. al., Science, May 15, 1998, vol 280, pp. 1077-1081.

More specifically, publication 1) is an example of a technology that uses mass spectrometry, specifically MALDITOF, for genotyping. Such a technology is a nonlimiting concrete example of a technology that is described in the application and is used by apparatus versions of the invention. And this concrete example allows for interpretation of means plus function language.

Publication 2) is an example of a technology that uses oligonucleotides for gentotyping, specifically high-density DNA arrays. Such a technology is a nonlimiting concrete example of a technology that is described in the application and is used by apparatus versions of the invention. A similar technology is described in publication 8), specifically genotyping chips. And each of these nonlimiting concrete example allows for interpretation of means plus function language.

Other examples of technologies that use oligonucleotides in various ways for genotyping, for example using PCR or other kinds of hybridization reactions, are described in the other publications enclosed herewith. Each of these technologies is a nonlimiting concrete example of a technology that is described in the application and is used by apparatus versions of the invention. And each of these nonlimiting concrete examples allows for interpretation of means plus function language.

There are other similar publications cited (and incorporated by reference) in the application that are not enclosed herewith. And this discussion is not necessarily exhaustive. No technology cited herein is admitted to being prior art with respect to the invention by its mention or discussion in this submission.

Sincerely,

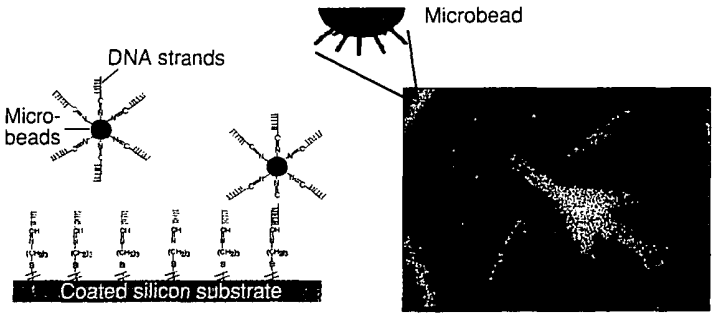
Robert O. McGinnis, Agent of Record, Reg. No. 44, 232

Publication 1)
With RCE Submission July 18, 2003 09/623,068

The San Diego researchers were looking for a way to help the right peg find its hole, and they settled on DNA. The chemical bases that make up DNA—cytosine, guanine, adenine, and thymine—will bind to each other only in particular pairings: C with G and A with T. Hence, a single strand made up of the bases ATTTGC will bind strongly with its complementary strand, TAAACG, and not with any other sequence. The researchers set out to exploit this selectivity by attaching short complementary strands of DNA to the pegs and substrate to help the devices find their correct positions.

In their first experiment, the team coated a substrate with a particular short strand of DNA. They then covered parts of the substrate with a mask and exposed it to ultraviolet light. The light chemically altered the DNA in exposed areas so that it could no longer bind to complementary strands. The researchers then coated some microbeads—which acted as dummy devices—with strands of DNA complementary to those on the substrate. When a fluid carrying the coated beads was splashed over the substrate, the beads successfully bound only to those areas that had not been exposed to UV light. One drawback of the technique is that it worked only for small devices, several hundred micrometers across, that would flow easily and not block other devices.

In a second experiment, designed to show that several varying kinds of "devices" could be deposited at once, the group used masks to deposit four different types of DNA strands onto a substrate and then attached complementary strands to four different fluorescent molecules. When the labeled molecules were splashed onto the substrate, the pattern of fluorescence showed that they had bound only to the appropriate regions of complementary DNA. In a real system, this would mean that four completely different types of devices could be attached to many selected sites on a chip.



Nature's glue. DNA strands bind beads and substrate together.

The researchers realize, however, that just providing the glue is not going to be enough. They are now looking for more active ways to guide the devices to their correct positions. One possibility is to add extra chemical groups to the DNA on the devices to give them an electric charge, then create electric fields on the substrate to attract the charged devices to "landing sites." The team is also investigating other techniques, such as creating currents in the fluid that would sweep the tiny devices to the right places.

An even bigger challenge will be creating an electrical connection between the devices and their host semiconductor. The team is looking at the possibility of putting the DNA glue on the top of devices and bonding them, upside down, onto a dummy substrate. Once all the devices are in position, the dummy could be flipped over and pressed down on the real substrate. The substrate might be coated with molten solder, which would add an electrical bond to the mass marriage.

–Sunny Bains

Sunny Bains is a science writer based in the San Francisco Bay area.

BIOTECHNOLOGY

Weighing DNA for Fast Genetic Diagnosis

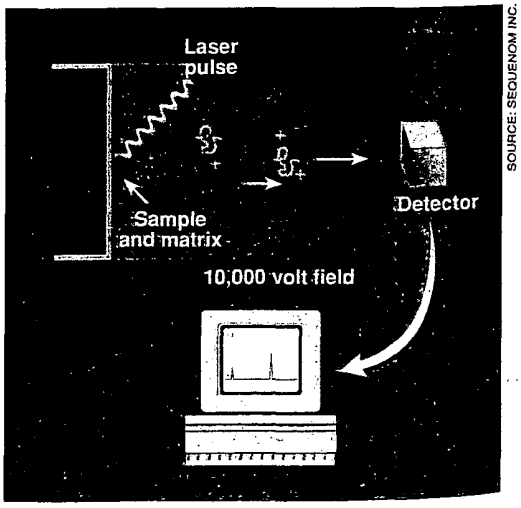
The modern doctor's little black bag, already overflowing with high-tech diagnostic devices, may soon have to make room for another advance. To diagnose a disease, judge future risks, or design a treatment, doctors will one day want to know which disease-related genes a patient carries. And they will want this diagnostic verdict to be as fast and accurate as a cholesterol or blood chemistry test today. As Charles Cantor, director of Boston University's Center for Advanced Biotechnology, puts it: "You need a detection system that can identify the gene sequences that you are looking for with high specificity, quickly, and in large volumes. The best analytical tool for doing this," he adds, "is mass spectrometry."

Borrowed from chemistry, this technology is a sharp departure from current methods, which identify a gene sequence by allowing it to bind to a matching probe, either on a gel or a chip. Instead, a mass spectrometer vaporizes the DNA and accelerates the molecules through a vacuum chamber with the help of an electric field. Tiny differences in the time it takes the DNA fragments to reach the detector reveal small differences in their mass, and hence their sequence.

The basic technique used for biomolecules is one with an unwieldy name, matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry, but a harmonious acronym, MALDI-TOF. It is now a decade old, but recent improvements have made it a hot commodity among companies hoping to commercialize DNA analysis. "With today's technology, MALDI-TOF can analyze hundreds of DNA samples ... in a matter of a few minutes," says Daniel P. Little, who directs mass-spectrometry development at Sequenom Inc., a San Diego–based company hoping to be generating diagnostic products within 6 months.

The standard way to distinguish different variants of a gene is to chop the DNA into fragments, separate them on a gel, and apply probes labeled with fluorescence or radioactivity, which bind to fragments with a particular sequence and light them up. But the process is slow and the gels can be hard to interpret. Newer techniques embed an array of different DNA probes on a single chip, allowing researchers to test for many gene variants at once. These so-called DNA chips can screen DNA quickly. But, as Cantor explains, the probes sometimes bind to sequences they don't completely match, which can limit the chips' accuracy.

Mass spectrometry may combine the DNA chip's speed with exquisite accuracy. The technique has long offered chemists a fast way to sort small molecules that vaporize naturally



All in the timing. A mass spectrometer sizes up DNA by vaporizing and ionizing it, accelerating the molecules, and recording their arrival times at a detector.

or can be coaxed into a vapor with bursts of energy from a laser or ion beam. But vaporizing large biomolecules while keeping them intact once seemed impossible. A decade ago, however, Franz Hillenkamp and colleagues at Westfälische Wilhelms University in Münster, Germany, found a way to do so with proteins: Cocrystallize them with certain small molecules, collectively called matrices. When a nanosecond laser pulse vaporizes the matrix, the resulting puff of material gently lifts the ionized biomolecule as well.

DNA was a tougher problem. But in 1993, Christopher Becker, then at SRI International in Palo Alto, California, and now at GeneTrace Systems in Menlo Park, California, found a simple matrix compound, 3-hydroxypicolinic acid, that worked with DNA sequences 20 to 25 bases long. By trial and error, MALDI practitioners have come up with several new matrices that work with DNA fragments as long as 100 bases.

The latest MALDI-TOF machines allow the cloud of matrix molecules to dissipate before applying an electric field. The field accelerates the charged DNA fragments toward a detector, and the differences in time of flight can reveal mass differences as small as 0.03%. If the DNA sequences from a gene have the same length—as they do if they have been produced by the polymerase chain reaction—any departure from the mass of the normal sequence reflects a mutation that has deleted or added bases or substituted others that have a different mass. "The results are an absolute indicator of the presence or absence of specific DNA sequences," says Sequenom's Little. MALDI-TOF can distinguish gene variants that differ by as little as a single base pair, and it can also analyze microsatellites—stretches of two-, three-, or four-nucleotide repeats often used as markers for locating disease-causing genes.

Besides offering unmatched precision, MALDI-TOF is inherently fast. The DNA forms a vapor and flies to the detector in fractions of a second; even repeating the process several times with the same sample to boost the sensitivity takes as little as 2 seconds. By preparing the samples in a grid and having the laser scan each spot in turn, a MALDI-TOF instrument can analyze 100 samples or more in a matter of minutes.

The combination of speed and accuracy could give the technique a role in genome sequencing as well as diagnosis. Standard, Sanger-type DNA sequencing generates many partial copies of a DNA sequence, each one starting at one end of the sequence and ending with a different one of the constituent bases. To determine the original sequence, biologists need to know the final base on each partial copy, together with the copy's length. Doing so now requires reading hundreds of bands on gels. But by sending the mixture through a mass spectrometer, biologists could quickly read off the fragments' lengths and—from the mass differences between successive fragments—the final base on each one. Investigators at both GeneTrace and Sequenom have published sequences determined with MALDI-TOF, the latest one, from Sequenom, appearing in the April *Nature Biotechnology*.

For practical gene sequencing, however, MALDI-TOF would have to work with DNA fragments much longer than the current 100-base capacity. Becker reportedly has discovered a new proprietary matrix that he expects will extend MALDI-TOF's reach to 1000-base sequences. "If you can really do upward of 1000 bases using this technique, and if it is indeed faster and cheaper, then this would be a big breakthrough for high-throughput sequencing," says Jeffrey Polish, who works in Mark Johnston's sequencing laboratory at Washington University in St. Louis.

In the meantime, the technology has no shortage of applications. Sequenom has shown, for example, that it can discriminate among 30 of the mutations that cause cystic fibrosis and pick up polymorphisms in the apolipoprotein E gene, which have been linked to familial hyperlipidemias, heart disease, and Alzheimer's disease. GeneTrace has developed a mass spectrometry–based system that can analyze which genes are being expressed in cells by identifying expressed sequence tags, short stretches of DNA copied from the messenger RNA made by active genes. Knowing which genes are active in a tissue can help pharmaceutical companies determine which ones are good drug targets.

With MALDI-TOF instruments running about $125,000 each—less than a standard clinical chemistry analyzer—these systems may also end up in large diagnostic labs. "Diagnostics at the level of the gene is something that we know is valuable, but is difficult, slow, and expensive today," says David Cooper, chief scientific officer at Nichols Institute Reference Laboratories, a division of Quest Diagnostics, one of the big 3 national reference laboratories. MALDI-TOF, he says, could be just the right medicine.

–Joseph Alper

Joseph Alper is a free-lance writer in Boulder, Colorado.

MATERIALS SCIENCE

Making a Bigger Chill With Magnets

LOS ANGELES—Refrigerator magnets are best known for holding shopping lists and old postcards onto refrigerator doors. But in a few years, much more powerful magnets could be the key to keeping food cold in so-called magnetocaloric refrigerators, which would be more energy efficient and less polluting than standard models. Now a new class of magnetocaloric materials, announced here last week at a meeting of the American Physical Society, could make these magnetic refrigerators more practical and versatile.

The magnetocaloric effect works when strong magnetic fields align quantum-mechanical "spins" of electrons within atoms. This transition reduces one aspect of the randomness, or entropy, of the atoms. But according to laws of thermodynamics, some other aspect of randomness has to increase in compensation, so the atoms increase the randomness of their velocities—vibrating and heating up. Once this heat is carried away by a coolant such as water, the field is removed and the effect works in reverse, chilling the material and cooling a refrigerator. To date, the peak performance has been with the element gadolinium.

By adding various amounts of silicon and germanium to gadolinium's crystal lattice, Vitalij Pecharsky and Karl Gschneidner of the Ames Laboratory at Iowa State University discovered a new class of materials that can chill two to six times further in a single magnetic cycle, meaning that the refrigerators could operate with weaker magnetic fields or less material. Depending on the germanium-to-silicon ratio, the new materials also operate from about room temperature all the way down to –253 degrees Celsius. The cold end of the range would allow magnetocaloric freezers to liquefy hydrogen or natural gas for use in clean-burning power plants or future automobiles.

To come up with the new compounds, the team followed up on hints that magnetocaloric materials containing gadolinium and either silicon or germanium—but not both—prefer a different range of temperatures than gadolinium alone. "We're not trying to come up with exotic new compounds out of the pure blue sky," says Gschneidner. The surprise, he says, was that the magnetocaloric effect turned out to be far larger when both germanium and silicon were added to the material.

"These new materials give you a lot more flexibility in designing magnetocaloric [refrigerators]," says Carl Zimm, a senior scientist in magnetic refrigeration at Astronautics Corporation of America in Madison, Wisconsin. The team is still working on making enough of the material to try it out in Zimm's prototype gadolinium-based refrigerator, which has been running for about a year. The test should take place "within a couple of months," says Gschneidner.

–James Glanz

Publication 2)
with RCE Submission July 18, 2003 09/623,068

D–0.1 M KCl, Tat-SF/pp140 was eluted with increasing salt concentrations and was detected mostly in 0.2 to 0.4 M KCl fractions. These fractions were pooled, dialyzed against buffer D–0.1 M KCl, and loaded onto a glutathione Sepharose (Pharmacia) column containing GST-Tat fusion proteins. After the column was washed with buffer D–0.4 M KCl, Tat-SF/pp140 was eluted from the column with buffer D containing 1.4 M KCl. The estimated overall purification after these steps was ~3000-fold. In the experiment shown in Fig. 3, the 0.2 to 0.4 M KCl heparin Sepharose fraction containing Tat-SF activity was subjected to fractionation through an Affi-Gel 10 matrix column (Bio-Rad) containing immobilized Tat. Tat-SF activity was eluted from the column with increasing salt concentrations. The 0.6 M KCl fraction was analyzed as described in Fig. 3.

10. T. O'Brien, S. Hardin, A. Greenleaf, J. T. Lis, *Nature* **370**, 75 (1994); M. E. Dahmus, *Biochim. Biophys. Acta.* **1261**, 171 (1995).
11. A. P. Rice and F. Carlotti, *J. Virol.* **64**, 1864 (1990).
12. The Tat-SF/pp140 fraction eluted from the GST-Tat column was subjected to SDS–polyacrylamide gel electrophoresis (PAGE), and the pp140 polypeptide was blotted onto a nitrocellulose membrane. Approximately 15 μg of pp140 were recovered from the membrane and subjected to digestion with lys-C. Six major peptides were obtained and microsequenced. One of the peptides (KMNAQETATGMAFEEPIDE) was contained in the sequence of EST60354 in the Washington University–Merck EST database. An Xho I–Eco RI fragment corresponding to the COOH-terminus of the Tat-SF1 gene and its 3′ untranslated region was labeled and used as a probe to screen a λZipLox (Gibco BRL) cDNA library prepared from human HL60 cells. Complementary DNAs were recovered from seven independent plaques in the autonomously replicating plasmid pZL1 as instructed by the manufacturer (Gibco BRL). The largest cDNA clone containing the full-length Tat-SF1 gene was named pZL-Tat-SF1-4b and was sequenced by dideoxy-DNA sequencing with T7 DNA polymerase.
13. D. R. Marshak and D. Carroll, *Methods Enzymol.* **200**, 134 (1991).
14. D. J. Kenan, C. C. Query, J. D. Keene, *Trends Biochem. Sci.* **16**, 214 (1991).
15. O. Delattre *et al.*, *Nature* **359**, 162 (1992); P. H. Sorensen *et al.*, *Nature Genet.* **6**, 146 (1994).
16. A. Crozat, P. Aman, N. Mandahl, D. Ron, *Nature* **363**, 640 (1993); T. H. Rabbitts, A. Forster, R. Larson, P. Nathan, *Nature Genet.* **4**, 175 (1993).
17. M. Ladanyi, *Diagn. Mol. Pathol.* **4**, 162 (1995); T. H. Rabbitts, *Nature* **372**, 143 (1994).
18. S. E. Harper, Y. Qiu, P. A. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8536 (1996).
19. J. W. Lillie and M. R. Green, *Nature* **338**, 39 (1989).
20. H. Kato *et al.*, *Genes Dev.* **6**, 655 (1992); R. A. Marciniak and P. A. Sharp, *EMBO J.* **10**, 4189 (1991).
21. M. G. Izban and D. S. Luse, *Genes Dev.* **6**, 1342 (1992); D. Wang and D. K. Hawley, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 843 (1993).
22. E. Bengal, O. Flores, A. Krauskopf, D. Reinberg, Y. Aloni, *Mol. Cell. Biol.* **11**, 1195 (1991); J. Greenblatt, J. R. Nodwell, S. W. Mason, *Nature* **364**, 401 (1993).
23. C. H. Herrmann and A. P. Rice, *J. Virol.* **69**, 1612 (1995).
24. N. A. McMillan *et al.*, *Virology* **213**, 413 (1995).
25. W. A. May *et al.*, *Mol. Cell. Biol.* **13**, 7393 (1993); H. Zinszner, R. Albalat, D. Ron, *Genes Dev.* **8**, 2513 (1994); D. D. Prasad, M. Ouchida, L. Lee, V. N. Rao, E. S. Reddy, *Oncogene* **9**, 3717 (1994).
26. P. J. Mitchell and R. Tjian, *Science* **245**, 371 (1989).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
28. M. A. Truett *et al.*, *DNA* **4**, 333 (1985).
29. H. E. Gendelman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9759 (1986).
30. L. S. Tiley, P. H. Brown, B. R. Cullen, *Virology* **178**, 560 (1990).
31. J. R. Neumann, C. A. Morency, K. O. Russian, *BioTechniques* **5**, 444 (1987).
32. We are grateful to B. Pepinsky and Biogen for providing pure HIV Tat protein and Tat mutant TatΔC; to J. Borrow [Massachusetts Institute of Technology (MIT) Center for Cancer Research] for human cDNA libraries; and to R. Cook (MIT Biopolymers Laboratory) for peptide sequencing. We thank K. Luo, J. Borrow, and H. Kawasaki for valuable advice and discussions; and B. Blencowe, K. Cepek, G. Jones, K. Luo, and C. Query for helpful comments on the manuscript. We also thank M. Siafaca for secretarial support. Supported by grants from the National Institutes of Health (GM34277 and AI32486) to P.A.S., and partially supported by a National Cancer Institute Center core grant (CA14051). Q.Z. was supported by a postdoctoral fellowship of The Jane Coffin Childs Memorial Fund for Medical Research.

19 June 1996; accepted 23 August 1996

Accessing Genetic Information with High-Density DNA Arrays

Mark Chee, Robert Yang, Earl Hubbell, Anthony Berno, Xiaohua C. Huang, David Stern, Jim Winkler, David J. Lockhart, Macdonald S. Morris, Stephen P. A. Fodor

Rapid access to genetic information is central to the revolution taking place in molecular genetics. The simultaneous analysis of the entire human mitochondrial genome is described here. DNA arrays containing up to 135,000 probes complementary to the 16.6-kilobase human mitochondrial genome were generated by light-directed chemical synthesis. A two-color labeling scheme was developed that allows simultaneous comparison of a polymorphic target to a reference DNA or RNA. Complete hybridization patterns were revealed in a matter of minutes. Sequence polymorphisms were detected with single-base resolution and unprecedented efficiency. The methods described are generic and can be used to address a variety of questions in molecular genetics including gene expression, genetic linkage, and genetic variability.

A central theme in modern genetics is the relation between genetic variability and phenotype. To understand genetic variation and its consequences on biological function, an enormous effort in comparative sequence analysis will need to be carried out. Conventional nucleic acid sequencing technologies make use of analytical separation techniques to resolve sequence at the single nucleotide level (*1*, *2*). However, the effort required increases linearly with the amount of sequence. In contrast, biological systems read, store, and modify genetic information by molecular recognition (*3*). Because each DNA strand carries with it the capacity to recognize a uniquely complementary sequence through base pairing, the process of recognition, or hybridization, is highly parallel, as every nucleotide in a large sequence can in principle be queried at the same time. Thus, hybridization can be used to efficiently analyze large amounts of nucleotide sequence. In one proposal, sequences are analyzed by hybridization to a set of oligonucleotides representing all possible subsequences (*4*). A second approach, used here, is hybridization to an array of oligonucleotide probes designed to match specific sequences. In this way the most informative subset of probes is used. Implementation of these concepts relies on recently developed combinatorial technologies to generate any ordered array of a large number of oligonucleotide probes (*5*).

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051, USA.

The fundamentals of light-directed oligonucleotide array synthesis have been described (*5*, *6*). Any probe can be synthesized at any discrete, specified location in the array, and any set of probes composed of the four nucleotides can be synthesized in a maximum of $4N$ cycles, where N is the length of the longest probe in the array. For example, the entire set of $\sim 10^{12}$ 20-nucleotide oligomer probes, or any desired subset, can be synthesized in only 80 coupling cycles. The number of different probes that can be synthesized is limited only by the physical size of the array and the achievable lithographic resolution (*7*).

An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Many different arrays can be designed for these purposes. One such design, termed a 4L tiled array, is depicted in Fig. 1A. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. By this approach, a nucleic acid target of length L can be scanned for mutations with a tiled array containing $4L$ probes. For example, to query the 16,569 base pairs (bp) of human mitochondrial DNA (mtDNA), only 66,276 probes of the possible $\sim 10^9$ 15-nucleotide oligomers need to be used.

The use of a tiled array of probes to read a target sequence is illustrated in Fig. 1C. A tiled array of 15-nucleotide oligomers varied

at position 7 from the 3′ end ($P^{15,7}$) was designed and synthesized for mt1, a cloned sequence containing 1311 bp spanning the control region of mtDNA (*8–11*). The upper panel of Fig. 1C shows a portion of the fluorescence image of an array hybridized with fluorescein-labeled mt1 RNA (*12*). The base sequence can be read by comparing the intensities of the four probes within each column. For example, the column for position 16,493 consists of the four probes, 3′-TGACATAGGCTGTAG, 3′-TGACATCGGCTGTAG, 3′-TGACATGGGCTGTAG, and 3′-TGACATTGGCTGTAG. The probe with the strongest signal is the probe with the A substitution (A, 49 counts; C, 8 counts, G, 15 counts, and T, 8 counts, where the background is 2 counts), identifying the base at position 16,493 as U in the RNA transcript. Continuing the process, the sequence at each position can be read directly from the hybridization intensities.

The effect on the array hybridization pattern caused by a single base change in the target is illustrated in Fig. 1B, and the detection of a single-base polymorphism is shown in the lower panel of Fig. 1C. The target was mt2, which differs from mt1 in this region by a T-to-C transition at position 16,493. Accordingly, the probe with the G substitution (third row) displays the strongest signal. Because the tiled array was designed to complement mt1, the hybridization intensities of neighboring probes that overlap position 16,493 are also affected by the change in target sequence. The hybridization signals of 15 probe sets of the 15-nucleotide oligomer tiled array are perturbed by a single base change in the target sequence. In the $P^{15,7}$ array, each probe querying the eight positions to the left and six positions to the right of the polymorphism contain at least one mismatch to the target. The result is a characteristic loss of signal or a "footprint" for the probes flanking a mutation position. Of the four probes querying each position, the loss of signal is greatest for the one designed to match mt1. We denote the subset of probes with zero mismatches to the reference sequence as P^0.

A comparison of P^0 hybridization signals from a target to those from a reference is ideally obtained by hybridizing both samples to the same array. We therefore developed a two-color labeling and detection scheme in which the reference is labeled with phycoerythrin (red), and the target with fluorescein (green) (*13*). By processing the reference and target together, experimental variability during the fragmentation, hybridization, washing, and detection steps is minimized or eliminated. In addition, during cohybridization of the reference and target, competition for binding sites results in a slight improvement in mismatch discrimination. Array hybridization is highly reproducible, and comparative analysis of data obtained from separate but identically synthesized arrays is also effective.

The two-color approach was tested by analyzing a 2.5-kb region of mtDNA that spans the $tRNA^{Glu}$, cytochrome b, $tRNA^{Thr}$, $tRNA^{Pro}$, control region, and $tRNA^{Phe}$ DNA sequences (*14*). A $P^{20,9}$ array (20-nucleotide oligomer probes varied at position 9 from the 3′ end) was designed to match the mt1 target (that is, P^0 sequence = mt1). The mt1 reference (red) and a polymorphic target sample (green) were pooled and hybridized simultaneously to the array. Differences between the target and reference sequences were identified by comparing the scaled red and green P^0 hybridization intensities (*15*). The marked decrease in target hybridization intensity, over a span of ~20 nucleotides, is shown for a single-base polymorphism at position 16,223 (Fig. 2A). The footprint is enlarged when two polymorphisms occur in close proximity (within ~20 nucleotides) (Fig. 2B). When polymorphisms are clustered, the size of the footprint depends on the number of polymorphisms and their separation (Fig. 2C).

To read polymorphisms accurately, we developed an algorithm that addresses the issue of multiple mismatches. The algorithm performs base identification but also flags regions of ambiguity caused by multiple mismatches. These regions are easily identified by the presence of a large footprint (Fig. 2, B and C) or by two or more bases identified as differing from P^0 within the span of a single probe. Discrepancies between base identifications and footprint patterns are also flagged for further analysis (for example, a P^0 footprint in which no polymorphism is identified; such a pattern is typical of a deletion). Thus, base identifications are valid only for unflagged regions. In flagged regions, the presence of sequence differences is detected, but no attempt is made to identify the sequence without further analysis.

Sequence analysis was carried out on the 2.5-kb target from 12 samples. A total of 30,582 bp containing 180 substitutions relative to mt1 were analyzed. Ninety-eight per-

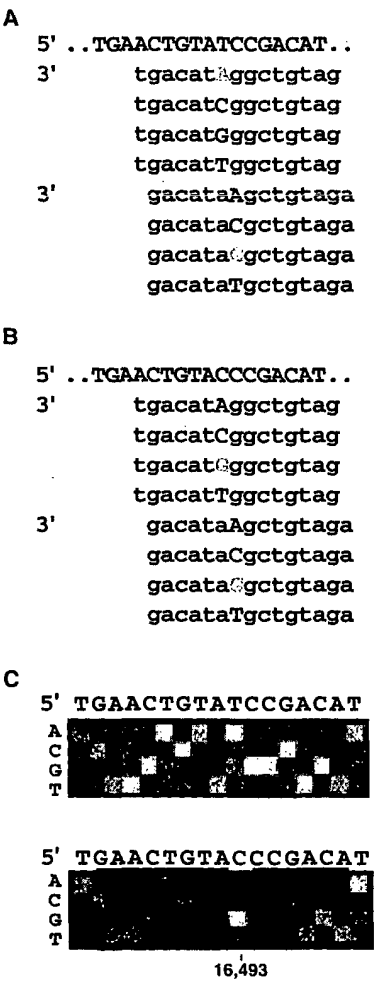


Fig. 1. (**A**) Design of a 4L tiled array. Each position in the target sequence (uppercase letters) is queried by a set of four probes on the chip (lowercase letters), identical except at a single position, termed the substitution position, which is either A, C, G, or T (blue indicates complementarity, red a mismatch). Two sets of probes are shown, querying adjacent positions in the target. (**B**) Effect of a change in the target sequence. The probes are the same as in (A), but the target now contains a single-base substitution (base C, shown in green). The probe set querying the changed base still has a perfect match (the G probe). However, probes in adjacent sets that overlap the altered target position now have either one or two mismatches (red) instead of zero or one, because they were designed to match the target shown in (A). (**C**) Hybridization to a 4L tiled array and detection of a base change in the target. The array shown was designed to the mt1 sequence. (Top) hybridization to mt1. The substitution used in each row of probes is indicated to the left of the image. The target sequence can be read 5′ to 3′ from left to right as the complement of the substitution base with the brightest signal. With hybridization to mt2 (bottom), which differs from mt1 in this region by a T→C transition, the G probe at position 16,493 is now a perfect match, with the other three probes having single-base mismatches (A 5, C 3, G 37, T 4 counts). However, at flanking positions, the probes have either single- or double-base mismatches, because the mt2 transition now occurs away from the query position.

cent of the sequence was unambiguously assigned by a Bayesian base identification algorithm (*16*). Of this 98%, which contained both wild-type sequence and a high proportion of single-base footprints such as the example shown in Fig. 2A, 29,878 out of 29,879 bp were identified correctly (*17*). The remaining 2% of the sequence, which contained the multiple substitution footprints (such as those shown in Fig. 2, B and C), was flagged for further analysis. Of the 649 bp composing this 2%, 643 bp were located in or immediately adjacent to footprints (*18*). In all, 179 out of the 180 polymorphisms were unambiguously detected, 126 out of 127 were identified correctly in the unflagged regions, and 53 polymorphisms occuring in the flagged regions were detected as footprints. There were no unflagged false-positive base identifications, and only one false-positive footprint. These figures can be considered to be "worst case" estimates for the type of array and target used. The P^0 sequence represents a Caucasian haplotype, and our sample set included eight African samples having a large number of clustered differences to P^0. Furthermore, the variation in the hypervariable part of the control region is much higher than for the rest of the mitochondrial genome and for nuclear genes in general (Fig. 2 shows comparisons to African samples in this region).

The determination of a complete human mitochondrial DNA sequence more than 15 years ago has had a tremendous influence on studies of human origins and evolution and the role of mutations in degenerative diseases (*8, 10, 19*). Because of the cost and difficulty of conventional sequence analysis, most subsequent sequencing studies have focused only on two small hypervariable regions totaling ~600 bp (*9*). However, access to the entire genome is required for a full understanding of the governing genetics. We therefore designed a $P^{25,13}$ tiling array for the mitochondrial genome. The array contains a total of 136,528 synthesis cells, each ~35 μm by 35 μm in size (Fig. 3). In addition to a 4L tiling across the genome, the array contains a set of probes representing a single-base deletion at every position across the genome and sets of probes designed to match a range of specific mtDNA haplotypes. Using long-range polymerase chain reaction, we amplified the 16.6-kb mtDNA directly from genomic DNA samples (*20*). Labeled RNA targets were prepared by in vitro transcription and hybridized to the array. Genomic hybridization patterns were imaged in less than 10 min by a high-resolution confocal scanner (*21*).

The hybridization pattern of a 16.6-kb target to the mitochondrial genome chip is shown in Fig. 3. Although there are some regions of low intensity, most of the 25-nucleotide oligomer array hybridized efficiently: Simply by identifying the highest intensity in each column of four substitution probes, 99.0% of the mt3 sequence could be read correctly (P^0 sequence = mt3). The array was used to successfully detect three disease-causing mutations in a mtDNA sample from a patient with Leber's hereditary optic neuropathy (*22, 23*) (Fig. 3C). In addition, we detected a total of seven errors and new polymorphisms from previously unsequenced regions.

We then hybridized 10 genomes from African individuals to the array and unambiguously identified 505 polymorphisms. These were polymorphisms that could be clearly read and for which a confirmatory footprint was detected automatically. For the 10 samples, the 2.5-kb cytochrome b and control region sequences were known (*17*). No false positives were detected in the ~25 kb of sequence checked in this way. Additional clustered polymorphisms were detected by the presence of footprints but not read directly. A detailed analysis of the polymorphisms in these genomes, and others, will be presented elsewhere.

The throughput of a conventional gel-based sequencer, with an average read length of 400 nucleotides and 48 lanes that is run twice a day, might be two mitochondrial genomes a day at best. In contrast, the throughput of the nonoptimized system we describe is five chips per hour. Thus, 50 genomes can be read by hybridization in the time it takes to read two genomes conventionally. Furthermore, there are significant reductions in sample preparation requirements because the entire genome is labeled in a single reaction, so the cost is similar to that for a single sequencing reaction. Also, sequence reading at the level of data analysis is automated: The sequences can be read in a matter of minutes. No analytical separations or gel preparation is needed, which contributes to the speed of the experiment. Although the inability to read all possible sequences is a weakness of the 4L tiled array, it is not a major limitation, because in practice the small number of ambiguities can be checked by targeted conventional sequencing. In particular, highly repetitive sequences, such as long runs of a single base, are presently best analyzed with conventional technology. Finally, a clear advantage to the approach we describe is that it is highly scalable. The cost, effort, and time required to analyze the entire 16.6-kb mtDNA in a single experiment is virtually identical to that required to read 2.5 kb. This provides a clear path to further orders-of-magnitude improvements in efficiency.

High-density oligonucleotide arrays

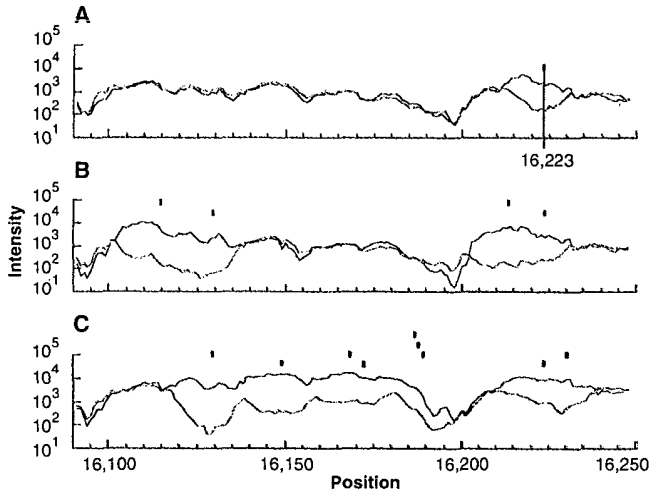


Fig. 2. Detection of base differences in a 2.5-kb region by comparison of scaled P^0 hybridization intensity patterns between a sample (green) and a reference (red) sequence. (**A**) Comparison of sequence ief007 to mt1. In the region shown, there is a single-base difference between the two sequences, located at position 16,223 (C in mt1, T in ief007). This results in a "footprint" spanning ~20 positions, 11 to the left and 8 to the right of position 16,223, in which the ief007 P^0 intensities are decreased by a factor of more than 10 on average relative to the mt1 intensities. The predicted footprint location is indicated by the gray bar, and the location of the polymorphism is shown by a vertical black line within the bar. The size of a footprint changes with probe length, and its relative position with substitution position (not shown). (**B**) Comparison of sequence ha001 to mt1. The ha001 target has four polymorphisms relative to mt1. The P^0 intensity pattern clearly shows two regions of difference between the targets. Each region contains two or more differences, because in both cases the footprints are longer than 20 positions and therefore are too extensive to be explained by a single-base difference. The effect of competition can be seen by comparing the mt1 intensities in the ief007 and ha001 experiments: The relative intensities of mt1 are greater in (B) where ha001 contains P^0 mismatches but ief007 does not. (**C**) The ha004 sequence has multiple differences to mt1, resulting in a complex pattern extending over most of the region shown. Thus, differences are clearly detected. Because hybridization intensities are extremely sequence-dependent, each of the mitochondrial sequences can also be identified simply by its hybridization pattern.

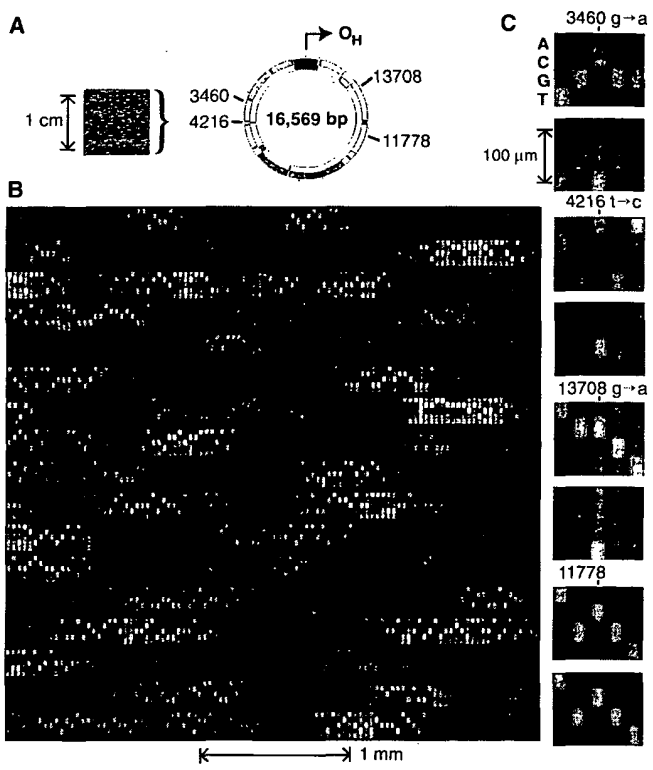


Fig. 3. Human mitochondrial genome on a chip. (**A**) An image of the array hybridized to 16.6 kb of mitochondrial target RNA (L strand). The 16,569-bp map of the genome is shown, and the H strand origin of replication (O_H), located in the control region, is indicated. (**B**) A portion of the hybridization pattern magnified. In each column there are five probes: A, C, G, T, and Δ, from top to bottom. The Δ probe has a single-base deletion instead of a substitution and hence is 24 instead of 25 bases in length. The scale is indicated by the bar beneath the image. Although there is considerable sequence-dependent intensity variation, most of the array can be read directly. The image was collected at a resolution of ~100 pixels per probe cell. (**C**) The ability of the array to detect and read single-base differences in a 16.6-kb sample is illustrated. Two different target sequences were hybridized in parallel to different chips. The hybridization patterns are compared for four different positions in the sequence. Only the $P^{25,13}$ probes are shown. The top panel of each pair shows the hybridization of the mt3 target, which matches the chip P^0 sequence at these positions. The lower panel shows the pattern generated by a sample from a patient with Leber's hereditary optic neuropathy (LHON). Three known pathogenic mutations, LHON3460, LHON4216, and LHON13708, are clearly detected. For comparison, the fourth panel in the set shows a region around position 11,778 that is identical in both samples.

provide the foundation for a powerful genetic analysis technology. The method can be used to characterize the spectrum of sequence variation in a population and can be applied to the analysis of many genes in parallel. In the case of human mtDNA, we simultaneously analyzed the control region, 13 protein coding genes, 22 tRNA genes, and 2 ribosomal RNA genes. The methods described here can be applied to other research areas in molecular genetics; for example, the ability to identify and sequence polymorphisms provides a basis for genetic mapping. The specificity of oligonucleotide hybridization and the scalability of the method suggests the possibility of a dedicated array that could be used to generate a high-resolution genetic map of an entire genome in a single experiment. Likewise, the concepts and techniques described here have been used to develop approaches for mRNA identification and the large-scale, parallel measurement of expression levels (*24*). Thus, the sequence of a gene, its spectrum of change in the population, its chromosomal location, and its dynamics of expression (all essential to a full understanding of function) can be determined with high-density probe arrays. The challenge now is to synthesize and read probe arrays at even higher density. For example, a 2 cm by 2 cm array, synthesized with probes occupying 1-μm synthesis sites in a 4L tiling, could query the entire coding content of the human genome, estimated at 100,000 genes.

REFERENCES AND NOTES

1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
2. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
3. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
4. W. Bains and G. C. Smith, *J. Theor. Biol.* **135**, 303 (1988); Y. P. Lysov *et al.*, *Dokl. Akad. Nank. SSSR* **303**, 1508 (1988); R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, *Genomics* **4**, 114 (1989); E. Southern, U. Maskos, R. Elder, *ibid.* **13**, 1008 (1992); see also R. B. Wallace *et al.*, *Nucleic Acids Res.* **6**, 3543 (1979).
5. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991).
6. A. C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994).
7. In the present format, we can routinely achieve a density of 409,600 synthesis sites in a 1.28 cm by 1.28 cm array. Each 20 μm by 20 μm site contains ~4 × 10^6 functional copies of a specific probe, which corresponds to a mean distance of about 100 Å between probes (M. O. Trulson, D. Stern, R. P. Rava, unpublished results).
8. S. Anderson *et al.*, *Nature* **290**, 457 (1981).
9. The control region of mtDNA is characterized by high amounts of sequence polymorphism concentrated in two hypervariable regions [B. D. Greenberg, J. E. Newbold, A. Sugino, *Gene* **21**, 33 (1983); C. F. Aquardo and B. D. Greenberg, *Genetics* **103**, 287 (1983)].
10. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* **106**, 479 (1984).
11. The mt1 and mt2 sequences were cloned from amplified genomic DNA extracted from hair roots [P. Gill, A. J. Jeffreys, D. J. Werrett, *Nature* **318**, 577 (1985); R. K. Saiki *et al.*, *Science* **239**, 487 (1988)]. The clones were sequenced conventionally (*1*). Cloning was performed only to provide a set of pure reference samples of known sequence. For templates for fluorescent labeling, DNA was reamplified from the clones with primers bearing bacteriophage T3 and T7 RNA polymerase promoter sequences (bold; mtDNA sequences uppercase): L15935-T3, 5′-ctcgg**aattaaccctcactaaagg**AAACCTTTTTCC-AAGGA and H667-T7, 5′-**taatacgactcactatagg**gagGGCTAGGACCAAACCTATT.
12. Labeled RNAs from the two complementary mtDNA strands [designated L and H (*8*)] were transcribed in separate reactions from a promoter-tagged polymerase chain reaction (PCR) product. Each 10-μl reaction contained 1.5 mM each of the triphosphate nucleotides ATP, CTP, GTP, and UTP; 0.24 mM fluorescein-12–CTP (Du Pont); 0.24 mM fluorescein-12–UTP (Boehringer Mannheim); ~1 to 5 nM (1.5 μl) crude unpurified 1.3-kb PCR product; and T3 or T7 RNA polymerase (1 U/μl) (Promega) in a reaction buffer supplied with the enzyme. The reaction was carried out at 37°C for 1 to 2 hours. RNA was fragmented to an average size of <100 nucleotides by adjusting the solution to 30 mM $MgCl_2$, by the addition of 1 M $MgCl_2$, and heating at 94°C for 40 min. Fragmentation improved the uniformity and specificity of hybridization (M. Chee *et al.*, data not shown). The extent of fragmentation is dependent on the magnesium ion concentration [J. W. Huff, K. S. Sastry, M. P. Gordon, W. E. C. Wacker, *Biochemistry* **3**, 501 (1964); J. J. Butzow and G. L. Eichorn, *Biopolymers* **3**, 95 (1965)]. Good hybridization results have been obtained with both DNA and RNA targets prepared with a variety of labeling schemes, including incorporation of fluorescent and biotinylated deoxynucleoside triphosphates by DNA polymerases, incorporation of dye-labeled primers during PCR, ligation of labeled oligonucleotides to fragmented RNA, and direct labeling by photo–cross-linking a psoralen derivative of biotin directly to fragmented nucleic acids (L. Wodicka, personal communication).
13. For two-color detection experiments, the reference and unknown samples were labeled with biotin and fluorescein, respectively, in separate transcription reactions. Reactions were carried out as described (*12*) except that each contained 1.25 mM of ATP, CTP, GTP, and UTP and 0.5 mM fluorescein-12–UTP or 0.25 mM biotin-16–UTP (Boehringer Mannheim). The two reactions were mixed in the ratio 1:5 (v/v) biotin:fluorescein and fragmented (*12*). Targets were diluted to a final concentration of ~100 to 1000 pM in 3M TMACl [W. B. Melchior Jr. and P. H. von Hippel, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 298 (1973)], 10 mM tris-HCl, pH 8.0, 1 mM EDTA, 0.005% Triton X-100, and 0.2 nM control oligonucleotide labeled at the 5′ end with fluorescein (5′-CTGAACGGTAG-CATCT TGAC). Samples were denatured at 95°C for 5 min, chilled on ice for 5 min, and equilibrated to 37°C. A volume of 180 μl of hybridization solution was then added to the flow cell [R. Lipshutz *et al.*, *Biotechniques* **19**, 442 (1995)] and the chip incubated at 37°C for 3 hours with rotation at 60 rpm. The chip was washed six times at room temperature with 6× SSPE (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, pH 7.4), 0.005% Triton X-100. Phycoerythrin-conjugated streptavidin (2 μg/ml in 6× SSPE, 0.005% Triton X-100) was added and incubation continued at room temperature for 5 min. The chip was washed again

and scanned at a resolution of ~74 pixels per probe cell. Two scans were collected: a fluorescein scan was obtained with a 515- to 545-nm band-pass filter, and a phycoerythrin scan with a 560-nm long-pass filter. Signals were separated to remove spectral overlap and average counts per cell determined.

14. Each 2.5-kb target sequence was PCR-amplified directly from genomic DNA with the primer pair L14675-T3 (5′-aattaaccctcactaaagggATTCTCGCACGGACTACAAC) and H667-T7 (*11*).
15. To scale the sample to the reference intensities, we constructed a histogram of the base 10 logarithm of the intensity ratios for each pair of probes. The histogram had a mesh size of 0.01 and was smoothed by replacing the value at each point with the average number of counts over a five-point window centered at that point. The highest value in the histogram was located, and the resulting intensity ratio was taken to be the most probable calibration coefficient.
16. Base identification was accomplished with a Bayesian classification algorithm based on variable kernel density estimation. The likelihood of each identification associated with a set of hybridization intensity values was computed by comparing an unknown set of probes to a set of example cases for which the correct base identification was known. The resulting four likelihoods were then normalized so that they summed to 1. Data from both strands were combined by averaging the values. If the most likely base identification had an average normalized likelihood greater than 0.6, it was called, otherwise the base was called as an ambiguity. The example set was derived from two different samples, ib013 and ief005, which have a total of 35 substitutions relative to mt1, of which 19 are shared with the 12 samples analyzed and 16 are not. Identification performance was not sensitive to the choice of examples.
17. To provide an independently determined reference sequence, each 2.5-kb PCR amplicon was sequenced on both strands by primer-directed fluorescent chain–terminator cycle sequencing with an ABI 373A DNA sequencer and assembled and manually edited with Sequencher 3.0. The analysis presented here assumes that the sequence amplified from genomic DNA is essentially clonal [R. J. Monnat and L. A. Loeb, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 2895 (1985)] and that its determination by gel-based methods is correct. A frequent length polymorphism at positions 303 to 309 was not detected by hybridization under the conditions used. It was excluded from analysis and is not part of the set of 180 polymorphisms discussed in the text. However, polymorphisms at this site have previously been differentiated by oligonucleotide hybridization [M. Stoneking, D. Hedgecock, R. G. Higuchi, L. Vigilant, H. A. Erlich, *Am. J. Hum. Genet.* **48**, 370 (1991)].
18. The P^0 intensity footprints were detected in the following way: The reference and sample intensities were normalized (*15*), and R_i the average of $\log(P^0_{reference}/P^0_{sample})$ over a window of five positions, centered at the base of interest, was calculated for each position in the sequence. Footprints were detected as regions having at least five contiguous positions with a reference or sample intensity at least 50 counts above background and an R value in the top 10th percentile for the experiment. At 205 polymorphic sites, where the sample was mismatched to P^0, the mean R value was 1.01, with a standard deviation of 0.57. At 35,333 nonpolymorphic sites (that is, where both reference and sample had a perfect match to P^0) the mean value was −0.05, with a standard deviation of 0.25.
19. R. L. Cann, M. Stoneking, A. C. Wilson, *Nature* **325**, 31 (1987); M. Zeviani *et al.*, *Am. J. Hum. Genet.* **47**, 904 (1990); D. C. Wallace, *Annu. Rev. Biochem.* **61**, 1175 (1992); S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, N. Takahata, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 532 (1995); T. Hutchin and G. Cortopassi, *ibid.*, p. 6892.
20. Long-range PCR amplification was carried out on genomic DNA with Perkin-Elmer GeneAmp XL PCR reagents according to the manufacturer's protocol. Primers were L14836-T3 (5′-aattaaccctcactaaagggATGAAACTTCGGCTCACTCCTTGGCG) and RH1066-T7 (5′-taatacgactcactatagggaTTTCATCATGCGGAGATGTTGGATGG), based on RH 1066 [S. Cheng, R. Higuchi, M. Stoneking, *Nature Genet.* **7**, 350 (1994)]. Each 100-μl reaction contained 0.2 μM concentration of each primer and ~10 to 50 ng of total genomic DNA. Transcription reactions were carried out in 10 μl with Ambion MAXIscript kit according to the manufacturer's protocol. The concentration of the 16.6-kb PCR template was ~2 nM, and the reaction contained Ambion 1× biotin-14-CTP/NTP mix and 0.2 mM biotin-16–UTP. Incubation was at 37°C for 2 hours. Fragmentation and hybridization were as described (*13*), except that 3.5 M TMACl and the biotin-labeled oligonucleotide 5′-CTGAACGGTAGCATCTTGAC were used in the hybridization buffer, which also contained fragmented baker's yeast RNA (100 μg/ml) (Sigma). Hybridization was carried out at 40°C for 4 hours.
21. A custom telecentric objective lens with a numerical aperture of 0.25 focuses 5 mW of 488-nm argon laser light to a 3-μm-diameter spot, which is scanned by a galvanometer mirror across a 14-mm field at 30 lines per second. Fluorescence collected by the objective is descanned by the galvanometer mirror, filtered by a dichroic beamsplitter (555 nm) and a band-pass filter (555 to 607 nm), focused onto a confocal pinhole, and detected by a photomultiplier. Photomultiplier output is digitized to 12 bits. A 4096 by 4096 pixel image is obtained in less than 3 min. Pixel size is 3.4 μm. The data from four sequential scans were summed to improve the signal-to-noise ratio.
22. M. D. Brown, A. S. Voljavec, M. T. Lott, I. MacDonald, D. C. Wallace, *FASEB J.* **6**, 2791 (1992).
23. Mitochondrial DNA populations can contain more than one sequence type, in a condition known as heteroplasmy. The LHON mutations shown in Fig. 3C were characterized as being homoplasmic by conventional sequencing and restriction endonuclease digestion (M. Brown, personal communication). In controlled mixing experiments, we have shown that sequences present at the level of 10% can easily be detected by hybridization (M. Chee and R. Yang, unpublished results; N. Shen, personal communication). The sensitivity of detection is sequence dependent. Importantly, hybridization can be used to detect heterozygous nuclear DNA sequences (J. Hacia *et al.*, in preparation).
24. D. J. Lockhart *et al.*, *Nature Biotech.*, in press.
25. We thank M. Brown and D. Wallace for the gift of the LHON sample and R. Ward for the 10 African samples, M. Trulson for assistance in two-color hybridization, P. Fiekowsky for image analysis, and P. Berg and E. Lander for comments on the manuscript. R. Davis contributed to the initial concepts in oligonucleotide tiling. We especially thank L. Stryer for his incessant and persistent encouragement. Supported in part by Human Genome grant 5RO1HG00813 from NIH (S.P.A.F.).

5 April 1996; accepted 26 July 1996

An Asymmetric Model for the Nucleosome: A Binding Site for Linker Histones Inside th DNA Gyres

Dmitry Pruss, Blaine Bartholomew, Jim Persinger, Jeffrey Hayes, Gina Arents, Evangelos N. Moudrianakis, Alan P. Wolffe*

Histone-DNA contacts within a nucleosome influence the function of trans-acting factors and the molecular machines required to activate the transcription process. The internal architecture of a positioned nucleosome has now been probed with the use of photoactivatable cross-linking reagents to determine the placement of histones along the DNA molecule. A model for the nucleosome is proposed in which the winged-helix domain of the linker histone is asymmetrically located inside the gyres of DNA that also wrap around the core histones. This domain extends the path of the protein superhelix to one side of the core particle.

The nucleosome has an active role in gene regulation. Mutations of the core histones have specific consequences for the transcription of particular genes (*1*). The specificity of these effects can be explained both by the positioning of histones with respect to DNA sequence (*2*) and the potential targeting of histone modifications to particular nucleosomes (*3*). Thus, an understanding of nucleosomal architecture is central to understanding the transcription process.

The nucleosome contains two molecules of each of the four core histones (H2A, H2B, H3, and H4), a single molecule of a linker histone (H1, H1°, or H5), and ~180 base pairs (bp) of DNA (*4*). In isolation, the core histones assemble into an octameric complex (*5*), whose structure has been determined at 3.1Å resolution (*6–8*). The exact path of DNA on the surface of the histone octamer, the position of the linker histone molecule within the nucleosome, and the path of linker DNA between adjacent nucleosomes (*9–11*) remain to be determined.

We used positioned nucleosomes containing the *Xenopus borealis* somatic 5S ribosomal RNA (rRNA) gene to examine

D. Pruss and A. P. Wolffe, Laboratory of Molecular Embryology, National Institute of Child Health and Human Development, National Institutes of Health, Building 6, Room B1A-13, Bethesda, MD 20892–2710, USA.
B. Bartholomew and J. Persinger, Department of Medical Biochemistry, Southern Illinois University at Carbondale, School of Medicine, Carbondale, IL 62901–4413, USA.
J. Hayes, Department of Biochemistry, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642, USA.
G. Arents and E. N. Moudrianakis, Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA.
*To whom correspondence should be addressed. E-mail: awlme@helix.nih.gov

Publication 3) 09/623,068
with RCE Submission
July 18, 2003

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 6230–6234, August 1989
Genetics

Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes

(polymerase chain reaction/"reverse dot blots"/nonradioactive detection/*HLA-DQA* locus/β-thalassemia)

RANDALL K. SAIKI*, P. SEAN WALSH*, COREY H. LEVENSON†, AND HENRY A. ERLICH*

Departments of *Human Genetics and †Chemistry, Cetus Corp., 1400 Fifty-Third Street, Emeryville, CA 94608

Communicated by Hamilton O. Smith, May 9, 1989 (received for review March 2, 1989)

ABSTRACT **The analysis of DNA for the presence of particular mutations or polymorphisms can be readily accomplished by differential hybridization with sequence-specific oligonucleotide probes. The *in vitro* DNA amplification technique, the polymerase chain reaction (PCR), has facilitated the use of these probes by greatly increasing the number of copies of target DNA in the sample prior to hybridization. In a conventional assay with immobilized PCR product and labeled oligonucleotide probes, each probe requires a separate hybridization. Here we describe a method by which one can simultaneously screen a sample for all known allelic variants at an amplified locus. In this format, the oligonucleotides are given homopolymer tails with terminal deoxyribonucleotidyltransferase, spotted onto a nylon membrane, and covalently bound by UV irradiation. Due to their long length, the tails are preferentially bound to the nylon, leaving the oligonucleotide probe free to hybridize. The target segment of the DNA sample to be tested is PCR-amplified with biotinylated primers and then hybridized to the membrane containing the immobilized oligonucleotides under stringent conditions. Hybridization is detected nonradioactively by binding of streptavidin-horseradish peroxidase to the biotinylated DNA, followed by a simple colorimetric reaction. This technique has been applied to *HLA-DQA* genotyping (six types) and to the detection of Mediterranean β-thalassemia mutations (nine alleles).**

Differential hybridization with sequence-specific oligonucleotide probes has become a widely used technique for the detection of genetic mutations and polymorphisms (1–5). When hybridized under the appropriate conditions, these synthetic DNA probes (usually 15–20 bases in length) will anneal to their complementary target sequences in the sample DNA only if they are perfectly matched. In most cases, the destabilizing effect of a single base-pair mismatch is sufficient to prevent the formation of a stable probe–target duplex (6). With an appropriate selection of oligonucleotide probes, the relevant genetic content of a DNA sample can be completely described.

This very powerful method of DNA analysis has been greatly simplified by the *in vitro* DNA-amplification technique, the polymerase chain reaction (PCR) (7–9). The PCR can selectively increase the number of copies of a particular DNA segment in a sample by many orders of magnitude. As a result of this 10^6- to 10^8-fold amplification, more convenient assays and nonradioactive detection methods have become possible (10–12). These PCR-based assays are usually done by amplifying the target segment in the sample to be tested, fixing the amplified DNA onto a series of nylon membranes, and hybridizing each membrane with one of the labeled oligonucleotide probes under stringent hybridization conditions. However, each probe must still be individually hybridized to the amplified DNA and the process can easily become difficult in a system where many different mutations or polymorphisms occur.

One approach to address this procedural difficulty is to "reverse" the DNAs: attach the oligonucleotides to the nylon support and hybridize the amplified sample to the membrane. Thus, in a single hybridization reaction, an entire series of sequences could be analyzed simultaneously. The strategy we adopted was to immobilize the oligonucleotides onto nylon filters by ultraviolet fixation. Exposure to UV light activates thymine bases in DNA, which then covalently couple to the primary amines present in nylon (13). It seemed unlikely, however, that short oligonucleotides could be directly attached to nylon in this manner and still retain their ability to discriminate at the level of a single base-pair mismatch. Consequently, the addition of a long deoxyribothymidine homopolymer tail, poly(dT), to the 3′ end of the oligonucleotide appeared promising for several reasons. First, the poly(dT) tail would be a larger target for UV crosslinking and should preferentially react with the nylon. Second, dTTP is very readily incorporated onto the 3′ ends of oligonucleotides by terminal deoxyribonucleotidyltransferase and would permit the synthesis of very long tails (14). (Deoxyribothymidine would also be the most efficiently incorporated base if a purely synthetic route were chosen.) Third, Collins and Hunsaker (15) had shown that the presence of a poly(dA) homopolymer tail, used to introduce multiple ^{35}S labels, did not affect the function of sequence-specific oligonucleotide probes.

We have used this technique to attach oligonucleotide probes specific for the six major *HLA-DQA* DNA types (16) and the eight most common Mediterranean β-thalassemia mutations (4) to nylon filters. The target segment of the DNA sample to be tested (either *HLA-DQA* or β-globin) was amplified by PCR with biotin-labeled primers to introduce a nonradioactive tag. Hybridization of the amplified product to the immobilized oligonucleotides and binding of streptavidin-horseradish peroxidase conjugate to the biotinylated primers were performed simultaneously. Detection was accomplished by a simple colorimetric reaction involving the enzymatic oxidation of a colorless chromogen that yielded a red color wherever hybridization occurred.

MATERIALS AND METHODS

Tailing of Oligonucleotides. Oligonucleotides were synthesized on a DNA synthesizer (model 8700, Biosearch) with β-cyanoethyl *N,N*-diisopropylphosphoramidite nucleosides (American Bionetics, Hayward, CA) by using protocols provided by the manufacturer. Oligonucleotide (200 pmol) was tailed in 100 μl of 100 mM potassium cacodylate/25 mM Tris·HCl/1 mM $CoCl_2$/0.2 mM dithiothreitol, pH 7.6 (17), with 5–160 nmol deoxyribonucleoside triphosphate (dTTP or

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: PCR, polymerase chain reaction.

dCTP) and 60 units (50 pmol) of terminal deoxyribonucleotidyltransferase (Ratliff Biochemicals, Los Alamos, NM) for 60 min at 37°C. Reactions were stopped by addition of 100 μl of 10 mM EDTA. The lengths of the homopolymer tails were controlled by limiting dTTP or dCTP. For example, a nominal tail length of 400 dT residues was obtained by using 80 nmol of dTTP in the above reaction.

Preparation of Filters. The tailed oligonucleotides were diluted into 100 μl of TE (10 mM Tris·HCl/0.1 mM EDTA, pH 8.0) and applied to a nylon membrane (Genetrans-45; Plasco, Woburn, MA) with a spotting manifold (BioDot; BioRad). The damp filters were then placed on TE-soaked paper pads in a UV light box (Stratalinker 1800; Stratagene) and irradiated at 254 nm. Dosage was controlled by the device's internal metering unit. The irradiated membranes were washed in 200 ml of 5× SSPE (1× SSPE is 180 mM NaCl/10 mM NaH_2PO_4/1 mM EDTA, pH 7.2) with 0.5% $NaDodSO_4$ for 30 min at 55°C to remove unbound oligonucleotides. If not used immediately, the filters were rinsed in water, air-dried, and stored at room temperature until needed.

Amplification of DNA. PCR amplification of genomic sequences was performed by a slight modification of previously described procedures (9). DNA (0.1–0.5 μg) was amplified in 100 μl containing 50 mM KCl, 10 mM Tris·HCl (pH 8.4), 1.5 mM $MgCl_2$, 10 μg of gelatin, 200 μM each dATP, dCTP, dGTP, and dTTP, 0.2 μM each biotinylated amplification primer, and 2.5 units of *Thermus aquaticus* (*Taq*) DNA polymerase (Perkin–Elmer/Cetus). The cycling reaction was done in a programmable heat block (DNA Thermal Cycler; Perkin–Elmer/Cetus) set to heat at 95°C for 15 sec (denature), cool at 55°C for 15 sec (anneal), and incubate at 72°C for 30 sec (extend) by the "Step-Cycle" program. After 30 repetitions, the samples were incubated an additional 5 min at 72°C. The primers contained a single molecule of biotin attached to the 5′ end of the oligonucleotides (described below).

Hybridization and Detection of Amplified DNA. Each filter with bound oligonucleotides was placed in 4 ml of hybridization solution containing 5× SSPE, 0.5% $NaDodSO_4$, and 400 ng of streptavidin-horseradish peroxidase conjugate (SeeQuence; Eastman Kodak). PCR-amplified DNA (20 μl) was denatured by addition of an equal volume of 400 mM NaOH/10 mM EDTA and added immediately to the hybridization solution, which was then incubated at 55°C for 30 min. (During this incubation, hybridization of PCR product to immobilized oligonucleotide and binding of streptavidin-horseradish peroxidase to biotin present in the PCR product occur simultaneously.) The filters were briefly rinsed twice in 2× SSPE/0.1% $NaDodSO_4$ at room temperature, washed once in 2× SSPE/0.1% $NaDodSO_4$ at 55°C for 10 min, and then briefly rinsed twice in 2× PBS (1× PBS is 137 mM NaCl/2.7 mM KCl/8 mM Na_2HPO_4/1.5 mM KH_2PO_4, pH 7.4) at room temperature. Color development was performed by incubating the filters in 25–50 ml of red leuco dye (Eastman Kodak) at room temperature for 5–10 min. Photographs were taken for permanent records.

Synthesis of Biotinylated Oligonucleotide Primers. Primary amino groups were introduced at the 5′ termini of the primers by a variation of published procedures (18, 19). In brief, tetraethylene glycol was converted to the monophthalimido derivative by reaction with phthalimide in the presence of triphenylphosphine and diisopropyl azodicarboxylate (20). The monophthalimide was converted to the corresponding β-cyanoethyl diisopropylamino phosphoramidite by standard protocols (21). The resulting phthalimido amidite was added to the 5′ ends of the oligonucleotides during the final cycle of automated DNA synthesis by using standard coupling conditions. During normal deprotection of the DNA (concentrated aqueous ammonia for 5 hr at 55°C), the phthalimido group was converted to a primary amine, which was subsequently acylated with an appropriate biotin active ester. NHS-LC-biotin (Pierce) was selected for its water solubility and lack of steric hindrance. The biotinylation was performed on crude, deprotected oligonucleotide, and the mixture was purified by a combination of gel filtration and reversed-phase HPLC. Additional details of this procedure will be published elsewhere (22).

RESULTS

Binding and Hybridization Efficiency of Tailed Oligonucleotides. The relative efficiencies with which synthetic oligonucleotides with homopolymer tails of various lengths were covalently bound to the nylon filter were measured as a function of UV exposure (Fig. 1 *Left*). Oligonucleotides with longer poly(dT) tails were more readily fixed to the membrane, and all attained their maximum values by 240 mJ/cm² of irradiation at 254 nm. In contrast, the $(dC)_{400}$-tailed oligonucleotide required more irradiation to crosslink to the nylon and was not comparable to the equivalent $(dT)_{400}$ construct even after 600 mJ/cm² exposure. This difference is consistent with the findings of Church and Gilbert (13) that suggested light-activated thymine bases bind more effectively to nylon than do cytosine bases. The untailed oligonucleotide was also retained by the membrane in a manner that roughly paralleled the poly(dC) product.

Efficient binding of oligonucleotides to the membrane, however, does not necessarily correlate with hybridization efficiency, and so hybridization efficiency as a function of UV dosage was determined in a separate experiment (Fig. 1 *Right*). These results show a distinct optimum of exposure that changes with the length of the poly(dT) tail and is more sharply pronounced for the longer tails. Additional experiments have shown the optimal dosages to be about 20 mJ/cm² for the $(dT)_{800}$ and 40 mJ/cm² for the $(dT)_{400}$ oligonucleotides (R.K.S., unpublished observations). The peak efficiencies of the $(dT)_{400}$ and $(dT)_{800}$ constructs are around 1% (45–50 fmol of radiolabeled probe annealed to ≈3.5 pmol of tailed oligonucleotide), which is similar to the value reported by Gamper *et al.* (23) for an oligonucleotide probe hybridized to nylon-bound plasmid DNA.

Comparison of the data in Fig. 1 *Left* and *Right* for 60 mJ/cm² irradiation indicates that oligonucleotides with longer tails hybridize more effectively than can be accounted for by the additional amounts bound to the filter. This suggests a spacer effect wherein the poly(dT) tails improve hybridization efficiency by increasing the distance between the nylon membrane and the terminal oligonucleotide probe. Besides possible UV damage to the DNA itself, additional exposure causes more of the tail to become attached to the membrane, thus reducing the average spacer length and decreasing hybridization efficiency. The markedly different hybridization profile of the poly(dC) oligonucleotide is compatible with this interpretation. Because cytosines react less efficiently with the filter, hybridization efficiency reaches a plateau where loss due to UV damage and tail shortening are compensated by the fixing of new molecules (see Fig. 1 *Left*). This characteristic of cytosine may make a poly(dC) tail desirable when UV irradiation cannot be carefully controlled. Under the stringent hybridization conditions used in this experiment, no signal was detected for the untailed oligonucleotide.

DNA Typing at the *HLA-DQA* Locus. The *HLA-DQA* test is derived from a PCR-based oligonucleotide typing system that partitions the polymorphic variants at the *DQA* locus into four major DNA types, *DQA1* to *DQA4*, and three *DQA1* subtypes, *DQA1.1* to *DQA1.3* (16). Four oligonucleotides specific for the major *DQA* types, four oligonucleotides that characterize the *DQA1* subtypes, and one control oligonucleotide that hybridizes to all allelic *DQA* sequences (Table

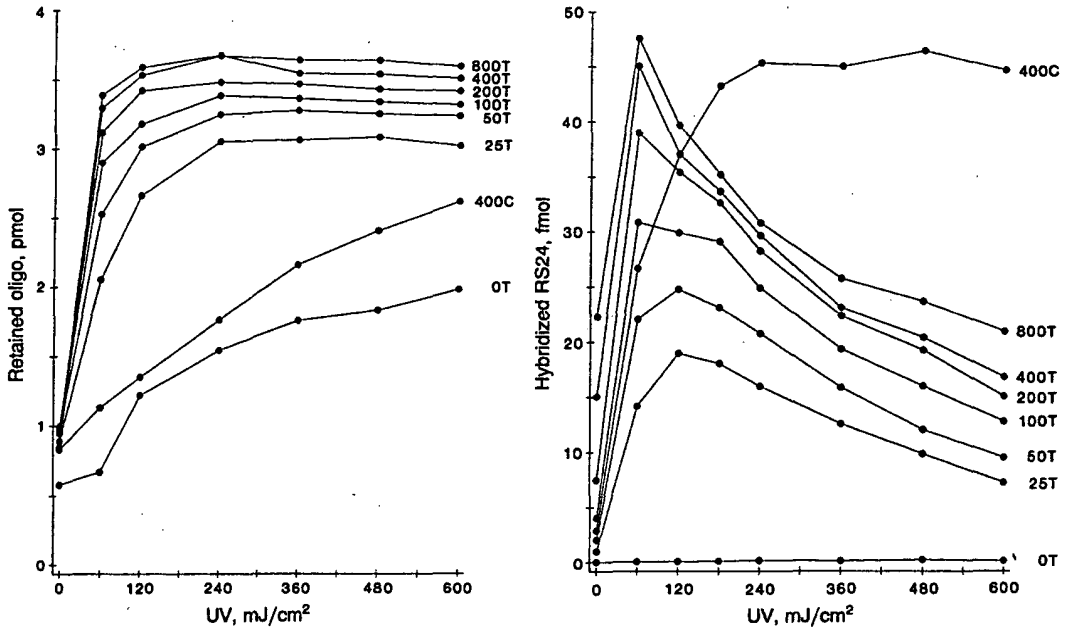


FIG. 1. Filter retention and hybridization efficiency of tailed oligonucleotides as a function of UV dosage and tail length. (*Left*) Filter retention. A 19-base oligonucleotide, 19A (5'-CTCCTGAGGAGAAGTCTGC-3'), was 5'-end-labeled with ^{32}P by using phage T4 polynucleotide kinase and [γ-^{32}P]ATP (10). Portions of the labeled oligonucleotide were given 3' homopolymer tails with terminal deoxyribonucleotidyltransferase and either dTTP or dCTP. The base compositions and lengths of the tails were as follows: $(dT)_0$, $(dT)_{25}$, $(dT)_{50}$, $(dT)_{100}$, $(dT)_{200}$, $(dT)_{400}$, $(dT)_{800}$, and $(dC)_{400}$. Four picomoles of each oligonucleotide was spotted onto nine duplicate filters, UV irradiated for various times, and washed to remove unbound oligonucleotides; each spot then was measured by scintillation counting to determine the amount crosslinked to the nylon. The values plotted are relative to an unirradiated, unwashed control filter (100% retention). (*Right*) Hybridization efficiency. Filters containing tailed, but unlabeled, 19A were prepared as described above and hybridized under sequence-specific conditions (see *Materials and Methods*) with a ^{32}P-labeled 40-base oligonucleotide, RS24 (5'-CCCACAGGGCAGTAACGGCAGACTTCTCCTCAGGAGTCAG-3'), complementary to 19A. The specific activity of the RS24 was 1.5 μCi/pmol (1 μCi = 37 kBq). Each spot was assayed by scintillation counting. The values plotted are fmol of RS24 hybridized to the membrane.

1) were given 400-base poly(dT) tails and spotted onto nylon filters. The sequence variation that defines the *DQA* types is localized within a relatively small "hypervariable" region of the second exon (24) that can be encompassed within a single 242-base-pair PCR amplification fragment. Biotinylated primers (Table 1) were used to amplify the *DQA* fragment from several genomic DNA samples: six homozygous cell lines and six heterozygous individuals. After hybridization of the amplified DNA to the membranes and color development, the *DQA* genotypes of these samples were readily apparent (Fig. 2).

Although most of the oligonucleotide probes are uniquely specific for one *DQA* type, two of the *DQA1* subtyping probes cross-hybridize to several DNA types. GH89 hybridizes to a sequence common to the *DQA1.2*, *DQA1.3*, and *DQA4* types, and the probe GH76 detects all *DQA* types except *DQA1.3*. (The latter is needed to distinguish *DQA1.2/1.3* heterozygotes from *DQA1.3/1.3* homozygotes.) The length and strand specificity of the oligonucleotides were empirically adjusted until their relative hybridization efficiencies and stringency requirements for allelic discrimination were approximately the same. (This was achieved by deter-

Table 1. Sequences of oligonucleotide primers and probes

Name*	Function	Sequence
RS151	*DQA* primer	b-GTGCTGCAGGTGTAAACTTGTACCAG†
RS152	*DQA* primer	b-CACGGATCCGGTAGCAGCGGTAGAGTTG†
RH54 (2)	All *DQA* types	CTACGTGGACCTGGAGAGGAAGGAGACTGCCTG
GH75 (4)	*DQA1* probe	CTCAGGCCACCGCCAGGCA
RH71 (4)	*DQA2* probe	TTCCACAGACTTAGATTTGAC
GH67 (4)	*DQA3* probe	TTCCGCAGATTTAGAAGAT
GH66 (4)	*DQA4* probe	TGTTTGCCTGTTCTCAGAC
GH88 (4)	*DQA1.1* probe	CGTAGAACTCCTCATCTCC
GH89 (4)	*DQA1.2, -1.3, -4*	GATGAGCAGTTCTACGTGG
GH77 (4)	*DQA1.3* probe	CTGGAGAAGAAGGAGAC
GH76 (4)	Not *DQA1.3*	GTCTCCTTCCTCTCCAG

Name*	Function	Sequence
RS151	β-Globin primer	b-ATCACTTAGACCTCACCCTG†
RS152	β-Globin primer	b-GACCTCCCACATTCCCTTTT†
RS187 (8)	Normal $\beta^{1\text{-}110}$	TAGACCAATAGGCAGAGAG
RS188 (8)	Mutant $\beta^{1\text{-}110}$	CTCTCTGCCTATTAGTCTA
RS87 (4)	Normal β^{39}	CCTTGGACCCAGAGGTTCT
RS89 (4)	Mutant β^{39}	AGAACCTCTAGGTCCAAGG
RS189 (0.33)	Normal $\beta^{1\text{-}1.6}$	CTTGATACCAACCTGCCCA‡
RS190 (0.33)	Mutant $\beta^{1\text{-}6}$	TGGGCAGGTTGGCATCAAG
RS191 (1)	Mutant $\beta^{1\text{-}1}$	TGGGCAGATTGGTATCAAG
RS192 (4)	Normal $\beta^{2\text{-}1}$	CCATAGACTCACCCTGAAG
RS193 (4)	Mutant $\beta^{2\text{-}1}$	CTTCAGGATGAGTCTATGG
RS201 (2)	Normal $\beta^{2\text{-}745}$	GCAGAATGGTAGCTGGATT
RS202 (2)	Mutant $\beta^{2\text{-}745}$	GCAGAATGGTACCTGGATT
RS196 (4)	Normal $\beta^{6\text{-}8}$	ACTCCTGAGGAGAAGTCTG‡
RS197 (4)	Mutant β^{6}	GACTCCTGGGAGAAGTCTG
RS198 (4)	Mutant β^{8}	TGACTCCTGAGGAGGTCTG

*Where applicable, the values in parentheses indicate the amount (pmol) of tailed oligonucleotide probe applied to the nylon membrane.
†b, Biotin covalently attached to 5' end.
‡These β-globin oligonucleotide probes each span two sites of potential β-thalassemia mutations and are specific for normal sequences at both positions.

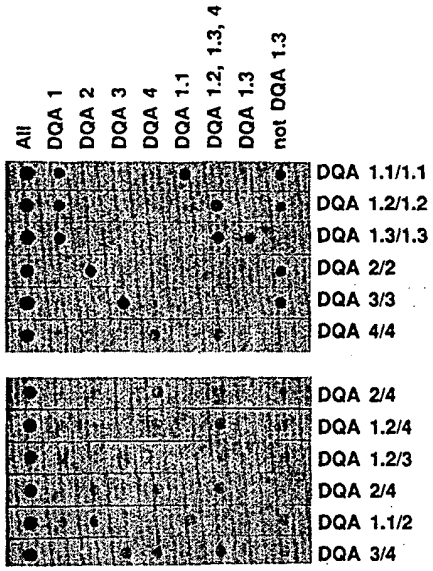


FIG. 2. DNA typing at the *HLA-DQA* locus. Each tailed oligonucleotide probe was spotted onto 12 duplicate membranes, irradiated at 40 mJ/cm^2, hybridized with amplified *DQA* sequences in genomic DNA samples, and treated for color development. The specificity of each immobilized oligonucleotide is given at the top, and the *DQA* genotype of each sample is noted at the right. The name, amount applied to the membrane, specificity, and sequence of each oligonucleotide are listed in Table 1.

mining the optimal hybridization conditions for each member of an initial set of probes, then shortening or lengthening each oligonucleotide until they all hybridized under equivalent conditions.) These eight probes produce a unique hybridization pattern for each of the 21 possible *DQA* diploid combinations.

Detection of β-Thalassemia Mutations. Although there are >54 characterized mutations of the β-globin gene that can give rise to β-thalassemia (25), each ethnic group in which this disease is prevalent has a limited number of common mutations (4, 26, 27). In Mediterranean populations, 8 mutations are responsible for >90% of the β-thalassemia alleles (4). Oligonucleotides were synthesized that are specific for each of these 8 mutations as well as their corresponding normal sequences (Table 1). The oligonucleotides were given $(dT)_{400}$ tails with terminal transferase and applied to membranes. Since the β-thalassemia mutations are distributed throughout the β-globin gene, biotinylated PCR primers that amplify the entire gene in a single 1780-base-pair fragment were used. (This amplification product encompasses *all* known β-thalassemia mutations, not only the predominant Mediterranean mutations examined here.) After hybridization and color development, the β-globin genotypes could be determined by noting the pattern of hybridization (Fig. 3).

Unlike the *DQA* typing system, two oligonucleotide probes are needed to analyze each mutation—one specific for the normal sequence and one specific for the mutant sequence—in order to differentiate normal/mutant heterozygous carriers from mutant/mutant homozygotes. A complicating factor in this analysis is caused by apparent secondary structure in various portions of the relatively long β-globin amplification product that interferes with oligonucleotide hybridization. The relatively high stringency needed to minimize this secondary structure requires the use of longer (e.g., 19-base) oligonucleotide probes. Because this constraint would not permit varying the length of the oligonucleotides to compensate for different hybridization efficiencies, the "balancing" of signal intensities was accomplished by adjusting the amount of each oligonucleotide applied to the membrane. This was done by applying various amounts of each oligonucleotide onto a membrane and then, after hybridization and color development, simply selecting the positive spots that had similar intensity.

DISCUSSION

These studies have demonstrated the feasibility of immobilizing sequence-specific probes onto nylon membranes and hybridizing PCR-amplified, biotin-labeled genomic fragments to the filters to determine the genetic content of the DNA sample. We have applied this method to *HLA-DQA* genotyping and to the detection of β-thalassemia mutations. Although the number of probes used in the two tests were modest (9 for *DQA* and 14 for β-thalassemia), expanding the analyses to include even more oligonucleotides should not be difficult.

The recently described technique of simultaneous amplification of several DNA fragments, "multiplex" PCR (28), should readily permit the concurrent analysis of multiple genetic loci. Using the immobilized-probe format, we have been able to simultaneously amplify and type at three loci: the *Hin*dIII polymorphism in the $^G\gamma$-globin gene (29), the *Ava* II

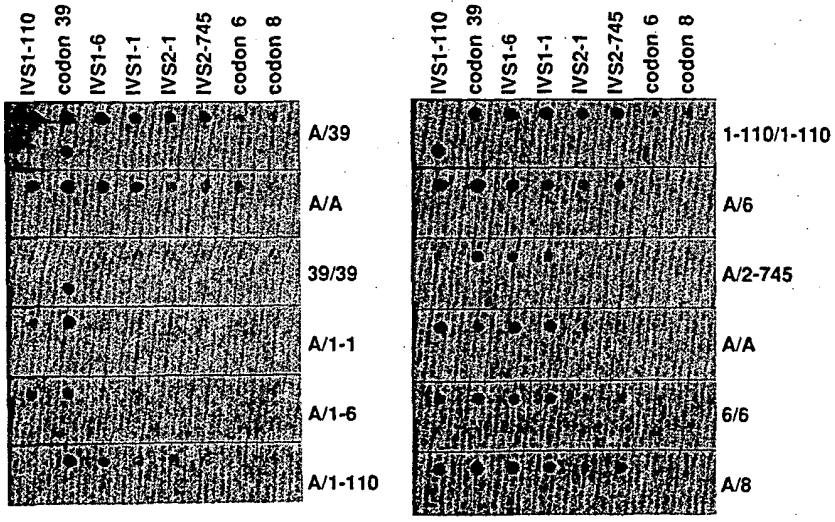


FIG. 3. Detection of β-thalassemia mutations. Various amounts of each tailed oligonucleotide probe were applied to 12 duplicate nylon filters, irradiated at 40 mJ/cm^2, hybridized with amplified β-globin sequences in genomic DNA samples, and treated for color development. The β-thalassemia locus that is detected by each immobilized oligonucleotide pair is given at the top of the filters. For each filter, the upper row contains the oligonucleotide probes that are specific for the normal sequence and the lower row contains the oligonucleotides specific for the mutant sequences. The β-globin genotype of each sample is noted at the right. The name, amount applied to the membrane, specificity, and sequence of each oligonucleotide are listed in Table 1. IVS, intervening sequence (intron).

polymorphism in the low density lipoprotein receptor gene (30), and the *HLA-DQA* gene (R.K.S., unpublished observations). Other genetic targets whose analysis would be simplified by this technique include the detection of somatic mutations in the *RAS* genes, where 6 loci and 66 possible alleles occur (31), some of the HLA class II β-chain genes, where as many as 25 alleles can be detected (T. Bugawan, S. Scharf, and H.A.E., unpublished observations), and β-thalassemia in Middle Eastern populations, where in addition to the endogenous mutations, Mediterranean and Asian Indian mutations are present at significant frequencies (H. Kazazian, personal communication). This format should also prove useful for the detection of infectious pathogens or for environmental surveys of microorganisms by immobilizing a panel of species-specific probes.

The ability to label probes and detect their hybridization without radioactivity is a convenient feature of PCR-based DNA tests and, perhaps more importantly, makes this type of analysis feasible in areas where radioactive labeling reagents are difficult to obtain. In this report, a biotin tag was introduced into the PCR products by means of 5'-biotinylated primers. An alternative labeling strategy based on the incorporation of biotinylated dUTP (32) has also been tried and shown to be very effective (R.K.S., unpublished observations).

One of the prerequisites of this analytical method is that all of the bound oligonucleotides must be sequence-specific under the same hybridization conditions. If necessary, this requirement can probably be met either by adjusting the length, position, and strand specificity of the probes, as was done for the *HLA-DQA* assay, or by varying the amount applied to the membrane, as was done for the β-thalassemia assay. The presence of tetramethylammonium chloride in the hybridization buffer can also serve to minimize the differences among immobilized oligonucleotides caused by varying base compositions (ref. 33; T. Bugawan, personal communication).

Although it may entail some initial effort, the end result is a simple, robust, and potentially automatable system that can be completed (amplification, hybridization, and color development) in 3–4 hr. "Reverse dot blots" should be particularly valuable for assays where the number of potential sequence variations exceeds the number of samples to be tested. Even in situations where the number of samples and probes are approximately equal, the immobilized-probe format may be preferable since many filters can be prepared at one time and stored until needed. To date, this typing system has been used to determine the *HLA-DQA* genotype of >300 unknown samples in forensic and disease-susceptibility studies.

We thank R. Higuchi and S. Scharf for helpful suggestions, L. Goda, D. Spasic, and C.-A. Chang for synthesis of oligonucleotides, C. Perez for advice on terminal transferase tailing reaction, C. Dowling and H. Kazazian (Johns Hopkins) for sequences of β-globin PCR primers and β-thalassemia genomic DNA samples, S. Warren and J. Findlay (Eastman Kodak) for red leuco dye suspension, and T. White, D. Gelfand, and H. Kazazian for critical review of the manuscript.

1. Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L. & Wallace, R. B. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 278–282.
2. Pirastu, M., Kan, Y. W., Cao, A., Conner, B. J., Teplitz, R. L. & Wallace, R. B. (1983) *N. Engl. J. Med.* **309**, 284–287.
3. Bos, J. L., Verlaan-de Vries, M., Jansen, A. M., Veeneman, G. H. & van Boom, J. H. (1984) *Nucleic Acids Res.* **12**, 9155–9163.
4. Kazazian, H. H., Jr., Orkin, S. H., Markham, A. F., Chapman, C. R., Youssoufian, H. & Waber, P. G. (1984) *Nature (London)* **310**, 152–154.
5. Kidd, V. J., Golbus, M. S., Wallace, R. B., Itakura, K. & Woo, S. L. (1984) *N. Engl. J. Med.* **310**, 639–642.
6. Ikuta, S., Takagi, K., Wallace, R. B. & Itakura, K. (1987) *Nucleic Acids Res.* **15**, 797–811.
7. Mullis, K. B. & Faloona, F. A. (1987) *Methods Enzymol.* **155**, 335–350.
8. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350–1354.
9. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
10. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1986) *Nature (London)* **324**, 163–166.
11. Bugawan, T. L., Saiki, R. K., Levenson, C. H., Watson, R. W. & Erlich, H. A. (1988) *Bio/Technology* **6**, 943–947.
12. Saiki, R. K., Chang, C.-A., Levenson, C. H., Warren, T. C., Boehm, C. D., Kazazian, H. H., Jr., & Erlich, H. A. (1988) *N. Engl. J. Med.* **319**, 537–541.
13. Church, G. M. & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1991–1995.
14. Nelson, T. & Brutlag, D. (1979) *Methods Enzymol.* **68**, 41–50.
15. Collins, M. L. & Hunsaker, W. R. (1985) *Anal. Biochem.* **151**, 211–224.
16. Higuchi, R., von Beroldingen, C. H., Sensabaugh, G. F. & Erlich, H. A. (1988) *Nature (London)* **332**, 543–546.
17. Roychoudhury, R. & Wu, R. (1980) *Methods Enzymol.* **65**, 43–62.
18. Coull, J. M., Weith, H. L. & Bischoff, R. (1986) *Tetrahedron Lett.* **27**, 3991–3994.
19. Connolly, B. A. (1987) *Nucleic Acids Res.* **15**, 3131–3139.
20. Mitsunobu, O. (1981) *Synthesis*, 1–28.
21. Sinha, N. D., McManus, J. & Koster, H. (1984) *Nucleic Acids Res.* **12**, 4539–4557.
22. Levenson, C. H. & Chang, C.-A. (1989) in *PCR Protocols and Applications: A Laboratory Manual*, eds. Innis, M. A., Gelfand, D. H. & White, T. J. (Academic, New York), in press.
23. Gamper, H. B., Cimino, G. D., Isaacs, S. T., Ferguson, M. & Hearst, J. E. (1986) *Nucleic Acids Res.* **14**, 9943–9954.
24. Horn, G. T., Bugawan, T. L., Long, C. M. & Erlich, H. A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6012–6016.
25. Kazazian, H. H., Jr., & Boehm, C. D. (1988) *Blood* **72**, 1107–1116.
26. Kazazian, H. H., Jr., Orkin, S. H., Antonarakis, S. E., Sexton, J. P., Boehm, C. D., Goff, S. C. & Waber, P. G. (1984) *EMBO J.* **3**, 593–596.
27. Zhang, J. Z., Cai, S. P., He, X., Lin, H. X., Lin, H. J., Huang, Z. G., Chehab, F. F. & Kan, Y. W. (1988) *Hum. Genet.* **78**, 37–40.
28. Chamberlain, J. S., Gibbs, R. A., Ranier, J. E., Nguyen, P. N. & Caskey, C. T. (1988) *Nucleic Acids Res.* **16**, 11141–11156.
29. Jeffreys, A. J. (1979) *Cell* **18**, 1–10.
30. Hobbs, H. H., Esser, V. & Russell, D. W. (1987) *Nucleic Acids Res.* **15**, 379.
31. Verlaan-de Vries, M., Bogaard, M. E., van den Elst, H., van Boom, J. H., van der Eb, A. J. & Bos, J. L. (1986) *Gene* **50**, 313–320.
32. Lo, Y. M., Mehal, W. Z. & Fleming, K. A. (1988) *Nucleic Acids Res.* **16**, 8719–8720.
33. Wood, W. I., Gitschier, J., Lasky, L. A. & Lawn, R. M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1585–1588.

Publication 4) 09/623,068
With RCE Submission
July 18, 2003

Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 2757–2760, April 1989
Genetics

Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia

(genetic diseases/base-pair mismatch/DNA polymerase/oligodeoxyribonucleotide/polymerase chain reaction)

DAN Y. WU*, LUIS UGOZZOLI†, BIJAY K. PAL‡, AND R. BRUCE WALLACE*

*Department of Molecular Biochemistry, Beckman Research Institute of the City of Hope, Duarte, CA 91010; †Laboratorio di Immunogenetica, Istituto Nazionale per la Ricerca sul Cancro, Genoa, Italy; and ‡Department of Biological Sciences, California State Polytechnic University, Pomona, CA 91768

Communicated by Eugene Roberts, December 27, 1988 (received for review December 12, 1988)

ABSTRACT **A rapid nonradioactive approach to the diagnosis of sickle cell anemia is described based on an allele-specific polymerase chain reaction (ASPCR). This method allows direct detection of the normal or the sickle cell β-globin allele in genomic DNA without additional steps of probe hybridization, ligation, or restriction enzyme cleavage. Two allele-specific oligonucleotide primers, one specific for the sickle cell allele and one specific for the normal allele, together with another primer complementary to both alleles were used in the polymerase chain reaction with genomic DNA templates. The allele-specific primers differed from each other in their terminal 3′ nucleotide. Under the proper annealing temperature and polymerase chain reaction conditions, these primers only directed amplification on their complementary allele. In a single blind study of DNA samples from 12 individuals, this method correctly and unambiguously allowed for the determination of the genotypes with no false negatives or positives. If ASPCR is able to discriminate all allelic variation (both transition and transversion mutations), this method has the potential to be a powerful approach for genetic disease diagnosis, carrier screening, HLA typing, human gene mapping, forensics, and paternity testing.**

Sickle cell anemia is the prototype of a genetic disease caused by a single base-pair mutation, an A → T transversion in the sequence encoding codon 6 of the human β-globin gene. In homozygous sickle cell anemia, the substitution of a single amino acid (Glu → Val) in the β-globin subunit of hemoglobin results in a reduced solubility of the deoxyhemoglobin molecule and erythrocytes assume irregular shapes. The sickled erythrocytes become trapped in the microcirculation and cause damage to multiple organs.

Kan and Dozy (1) were the first to describe the diagnosis of sickle cell anemia in the DNA of affected individuals based on the linkage of the sickle cell allele to an *Hpa* I restriction fragment length polymorphism. Later, it was shown that the mutation itself affected the cleavage site of both *Dde* I and *Mst* II and could be detected directly by restriction enzyme cleavage (2, 3). Conner *et al.* (4) described a more general approach to the direct detection of single nucleotide variation by the use of allele-specific oligonucleotide hybridization. In this method, a short synthetic oligonucleotide probe specific for one allele only hybridizes to that allele and not to others under appropriate conditions.

All of the above approaches are technically challenging, require a reasonably large amount of DNA, and are not very rapid. The polymerase chain reaction (PCR) developed by Saiki *et al.* (5) provided a method to rapidly amplify small amounts of a particular target DNA. The amplified DNA could then be readily analyzed for the presence of DNA sequence variation (e.g., the sickle cell mutation) by allele-specific oligonucleotide hybridization (6), restriction enzyme cleavage (5, 7), ligation of oligonucleotide pairs (8, 9), or ligation amplification (10). PCR increased the speed of analysis and reduced the amount of DNA required for it but did not change the method of analysis of DNA sequence variation. In this paper, we investigated whether PCR could be done in an allele-specific manner such that the presence or absence of an amplified fragment provides direct determination of genotype.

PCR utilizes two oligonucleotide primers that hybridize to opposing strands of DNA at positions spanning a sequence of interest. A DNA polymerase [either the Klenow fragment of *Escherichia coli* DNA polymerase I (5) or *Thermus aquaticus* DNA polymerase (11)] is used for sequential rounds of template-dependent synthesis of the DNA sequence. Prior to the initiation of each new round, the DNA is denatured and fresh enzyme is added in the case of the *E. coli* enzyme. In this manner, exponential amplification of the target sequences is achieved. We reasoned that if the 3′ nucleotide of one of the primers formed a mismatched base pair with the template due to the existence of single nucleotide variation, amplification would take place with reduced efficiency. Specific primers would then direct amplification only from their homologous allele. After multiple rounds of amplification, the formation of an amplified fragment would indicate the presence of the allele in the initial DNA.

MATERIALS AND METHODS

Oligonucleotide Synthesis. Oligonucleotides were synthesized on an Applied Biosystems 380B DNA synthesizer by the phosphoramidite method. They were purified by electrophoresis on a urea/polyacrylamide gel followed by high-performance liquid chromatography as described (12).

Source and Isolation of Human DNA. All genomic DNA samples with the exception of the β-thalassemia DNA were isolated from the peripheral blood of appropriate individuals. The β-globin genotype of these individuals was previously determined by hybridization with allele-specific oligonucleotide probes (4) as well as by hemoglobin electrophoresis. Thalassemia major DNA was obtained from an Epstein–Barr virus-transformed lymphocyte cell line obtained from the National Institute of General Medical Sciences Human Genetic Mutant Cell Repository (Camden, NJ). Thalassemia DNA was isolated from the cultured cells. All DNA preparations were performed according to a modified Triton X-100 procedure followed by proteinase K and RNase A treatment (13). The average yield of genomic DNA was ≈25 μg per ml of blood.

PCR. Hβ14A (5′-CACCTGACTCCTGA) and BGP2 (5′-AATAGACCAATAGGCAGAG) at a concentration of 0.12 μM were used as the primer set for the amplification of the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PCR, polymerase chain reaction; ASPCR, allele-specific PCR.

normal β-globin gene (*a* primer set). Similarly, 0.12 μM HβI4S (5'-CACCTGACTCCTGT) and 0.12 μM BGP2 were used as the primer set for the amplification of the sickle cell gene (*s* primer set). Both primer sets directed the amplification of a 203-base-pair (bp) β-globin allele-specific fragment. As an internal positive control, all reaction mixtures contained an additional primer set for the human growth hormone gene comprised of 0.2 μM GHPCR1 (5'- TTCCCAACCATTCCCTTA) and 0.2 μM GHPCR2 (5'-GGATTTCTGTTGTGTTTC) (*hGH* primer set). GHPCR1 and GHPCR2 direct the amplification of a 422-bp fragment of the human growth hormone gene. All reactions were performed in a vol of 50 μl containing 50 mM KCl, 10 mM Tris·HCl (pH 8.3), 1.5 mM $MgCl_2$, 0.01% (wt/vol) gelatin, template DNA (0.5 μg/ml), and 0.1 mM each dATP, dCTP, dGTP, and TTP. Reactions were carried out for 25 cycles at an annealing temperature of 55°C for 2 min, a polymerization temperature of 72°C for 3 min, and a heat-denaturation temperature of 94°C for 1 min on a Perkin-Elmer Cetus DNA thermal cycler. At the end of the 25 rounds, the samples were held at 4°C in the thermal cycler until removed for analysis.

Analysis of the PCR Products. An aliquot (15 μl) from each of the completed PCR reactions was mixed with 5 μl of 5× Ficoll loading buffer (1× = 10 mM Tris·HCl, pH 7.5/1 mM EDTA/0.05% bromophenol blue/0.05% xylene cyanol/3% Ficoll) and subjected to electrophoresis in a 1.5% agarose gel. Electrophoresis was performed in 89 mM Tris·HCl/89 mM borate/2 mM EDTA buffer for 2 hr at 120 V. At the completion of electrophoresis, the gel was stained in ethidium bromide (1.0 μg/ml) for 15 min, destained in water for 10 min, and photographed by ultraviolet trans-illumination.

RESULTS

Experimental Design. The scheme describing allele-specific PCR (ASPCR) is shown in Fig. 1. Primer P1 is designed such that it is complementary to allele 1 but the 3'-terminal nucleotide forms a single base-pair mismatch with the DNA sequence of allele 2 (Fig. 1*B*, *). Under appropriate annealing temperature and PCR conditions, there is normal amplification of the P1–P3 fragment with DNA templates containing allele 1 (homo- or heterozygous), while there is little or no amplification from DNA templates containing allele 2. In a similar way, a primer (P2) could be designed that would allow the specific PCR amplification of allele 2 but not allele 1 DNA.

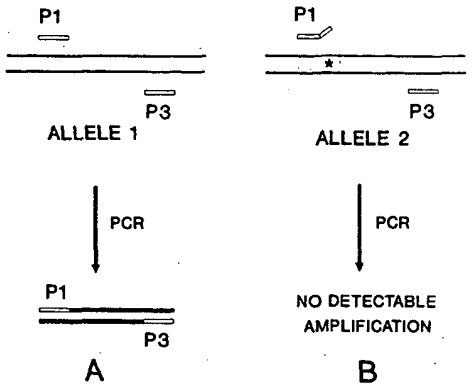


FIG. 1. Schematic representation of the ASPCR. P1 and P3, synthetic oligonucleotide primers that anneal to opposing strands of a single copy gene. P1 anneals to the region of a gene in the region of a DNA sequence variation such that its terminal 3' nucleotide base pairs with the polymorphic nucleotide of the template. P1 is completely complementary to allele 1 (*A*) but forms a single base-pair mismatch with allele 2 at the 3'-terminal position due to one or more nucleotide differences relative to allele 1 (*B*).

We designed two 14-nucleotide-long allele-specific primers, HβI4S and HβI4A, complementary to the 5' end of the sickle cell and normal β-globin genes, respectively. The oligonucleotide primers differ from each other by a single nucleotide at the 3' end, HβI4S having a 3' T and HβI4A having a 3' A corresponding to the base pair affected by the sickle cell mutation. The oligonucleotide primer BGP2 (7) complementary to the opposite strand 3' of the allele-specific primers was used as the second primer for PCR. The amplification product with these primer pairs was 203 bp. Also included in each reaction was a second pair of primers that directed the amplification of a 422-bp fragment of the human growth hormone gene. These primers were included as an internal positive control.

Discrimination Between the Normal and Sickle Cell Alleles. Genomic DNA was isolated from peripheral blood leukocytes of individuals of known β-globin genotypes (β^A/β^A, β^A/β^S, β^S/β^S). In addition, DNA was isolated from an Epstein–Barr virus-transformed cell line containing a homozygous deletion of the β-globin gene (β^{th}/β^{th}). DNA was subjected to 25 rounds of PCR using either the sickle cell-specific primer set (HβI4S and BGP2) or the normal gene-specific primer set (HβI4A and BGP2) using an annealing temperature of 55°C. The results are shown in Fig. 2*A*. It can be seen that a 203-bp fragment is observed using the sickle cell-specific primer set only with the β^A/β^S and β^S/β^S genomic DNA templates and not with the β^A/β^A genomic DNA templates. Conversely, the normal gene-specific primer set only gave rise to an amplification product with β^A/β^S and β^A/β^A genomic DNA templates. As expected, the thalassemia DNA did not give rise to a β-globin gene amplification product with either primer set. The internal growth hormone gene control gave rise to a 422-bp fragment in all samples, demonstrating that in no case was the absence of a globin-specific band due to a failure of the PCR.

In a single blind study, the DNA from 12 individuals with different β-globin genotypes was analyzed with the two primer sets. The results are shown in Fig. 2*B*. Individuals 1, 2, 3, and 5 are predicted to be β^A/β^A; individuals 6, 9, 10, and 11 are predicted to be β^S/β^S; and individuals 4, 7, 8, and 12 are predicted to be β^A/β^S. In each case, the genotype was correctly and unambiguously predicted from the pattern of fragment amplification (see legend to Fig. 2 for clinically diagnosed genotype).

DISCUSSION

The results presented above indicate the potential usefulness of ASPCR for sickle cell diagnosis. The method is rapid and the result is obtained without the use of radioactivity, since all that is required is to visualize the band on a gel with ethidium bromide staining. It should be possible to further improve the technique by elimination of the gel separation step. One strategy for this is shown in Fig. 3. As proposed recently by Yamane *et al.* (15), the two primers for the PCR could be labeled differently, one with biotin and one with a fluorescent group such as fluorescein or tetramethyl rhodamine. The product of the PCR could be captured on streptavidin-agarose and the presence of the amplified sequence could be detected with the fluorescence. In this case, if one allele-specific primer were labeled with one fluorescent group and the other were labeled with a different one, then the ASPCR could be done simultaneously.

In this study, we have used PCR primers that form either an A·A or a T·T mismatch. It is not clear that other mismatches will give equally effective discrimination. Since G·T mismatches are more stable than other mismatches (16), G·T should probably be avoided when designing primers.

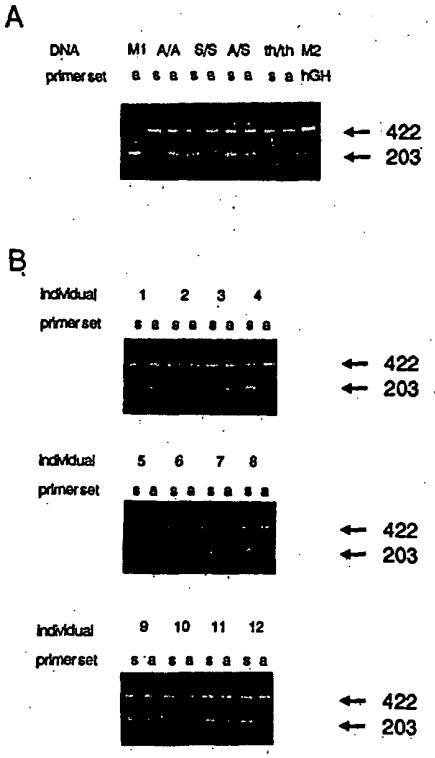


FIG. 2. (*A*) Identification of the normal (β^A) and the sickle cell (β^S) alleles by ASPCR. Normal (β^A/β^A), homozygous sickle cell (β^S/β^S), heterozygous sickle cell (β^A/β^S), and homozygous β-thalassemia (β^{th}/β^{th}) DNA samples (0.5 μg each) served as template using either the normal (*a* primer set) or the sickle cell (*s* primer set) for the ASPCRs. As an internal positive control, all reaction mixtures contained an additional primer set for the human growth hormone gene (*hGH* primer set) that directed the amplification of a 422-bp fragment of the human growth hormone gene. After amplification, 15 μl from each reaction mixture was subjected to electrophoresis in a 1.5% agarose gel for 2 hr at 120 V. Ethidium bromide staining of the agarose gel was used to detect PCR amplified fragments. Positive β-globin ASPCR can be identified by the presence of a 203-bp fragment using either the *a* or the *s* primer set reaction. As a marker for the globin-specific fragment, 0.3 μg of plasmid pHβ^A containing the normal human globin gene (β^A) was amplified with the *a* primer set alone (M1). As a marker for the growth hormone-specific fragment, 0.1 μg of plasmid pXGH5 containing a 3.8-kilobase fragment of the human growth hormone gene (14) was amplified with the growth hormone primer set (hGH) alone (M2). (*B*) A single blind trial using ASPCR to diagnose the β-globin genotype of genomic DNA samples. Genomic DNA samples from 12 individuals (4 each of normal, homozygous, and heterozygous sickle cell individuals) were randomly assigned numbers 1–12 by the hematology laboratory and blinded to the investigators. ASPCR was performed using both the normal (*a*) and the sickle cell-specific (*s*) primer sets as described above. Genotypes were identified as homozygous normal (β^A/β^A) if the single 203-bp fragment appears exclusively in the *a* primer set reaction, as homozygous sickle cell (β^S/β^S) if the 203-bp fragment appears only in the *s* primer set, or as heterozygous sickle cell trait (β^A/β^S) if the fragment appears in both reactions. The genotypes of these DNA samples were previously determined by hemoglobin electrophoresis (results not shown). The genotypes of the 12 individuals are as follows: 1, 2, 3, and 5, β^A/β^A; 6, 9, 10, and 11, β^A/β^S; 4, 7, 8, and 12, β^S/β^S.

This can be done by designing the primer so that it is complementary to the strand with which it forms an A·C mismatch. It may be possible to use a competition approach, as we have previously used to improve the discrimination provided by oligonucleotide hybridization probes (17). In this

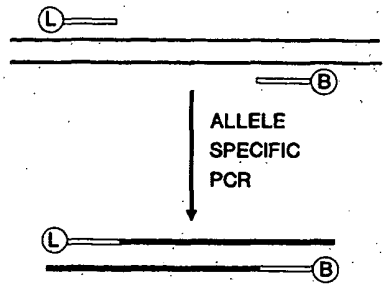


FIG. 3. Schematic representation of a dual labeling system suitable for the detection of the ASPCR products. One of the oligonucleotide primers is labeled at the 5′ end with a fluorescent group such as fluorescein or tetramethyl rhodamine (L) and the other primer is labeled with biotin (B). The ASPCR amplification product would therefore have the 5′ end labeled on both strands. The biotin is suitable for capturing the amplified fragment on a streptavidin-agarose column, while the fluorescent group is suitable for measuring the amount of fragment produced.

case, a competitive primer could be designed that was not able to prime, for example, by including in it a 3′ dideoxynucleotide or a 3′ ribonucleotide that has been oxidized. A mixture of a labeled allele-specific primer complementary to allele 1 plus an unlabeled priming-defective primer complementary to allele 2 should then allow the specific amplification of allele 1.

The ability of an oligonucleotide to prime on a DNA template is governed by two kinetic variables: the rate at which the annealed primer dissociates from the template before initiating polymerization (r_{off}) and the rate at which the DNA polymerase extends the primer (r_{pol}). Efficient priming in PCR should take place whenever $r_{pol} > r_{off}$, the addition of the first few nucleotides to the primer then greatly stabilizing the oligonucleotide–template complex and allowing continued extension of the primer. For a given primer r_{pol} is an intrinsic property of the polymerase. Studies with *E. coli* DNA polymerase I have suggested that this polymerase may be able to discriminate between primers that either do or do not form a mismatch with the template at the 3′-terminal nucleotide (18). In this case, r_{pol} for the mismatched primer was slower than r_{pol} for the perfectly matched primer. For the present study, we designed the allele-specific primers such that the allele-specific nucleotide in the template was complementary to the 3′-terminal nucleotide of the primer. In this way, the 3′ nucleotide of the primer specific for one allele would form a mismatch with the other allele. This design allows one to take advantage of the difference between r_{pol} of the perfectly matched and mismatched primers as well as to optimize primer concentration, priming temperature, primer length, and primer sequence, all of which will affect the difference in the r_{off} for the two allele-specific primers.

We reasoned that a set of conditions should exist such that $r_{pol} > r_{off}$ for the perfectly matched primer, while $r_{pol} < r_{off}$ for the mismatched primer. The results shown here clearly demonstrate this to be true. In our study, the allele-specific primers were 14 nucleotides long. We found (data not shown) that discrimination between the β^A and β^S alleles was not possible at low annealing temperatures (e.g., 44°C and 50°C). Presumably the short length of the oligonucleotides as well as the high annealing temperature combined to provide the discrimination.

Taq polymerase is well suited for using ASPCR for the discrimination of two alleles that differ by a single nucleotide because it lacks a 3′ → 5′ exonuclease activity (19). Such an activity would correct the mismatched base pair in the mismatched primer–template complex and then permit efficient priming with the one-nucleotide-shorter primer. Since

the specificity of the ASPCR is determined in the initial several cycles of PCR, the fact that the primer remains uncorreeted enhances the discrimination of the reaction. PCR is an exponential reaction; the yield of product is very dependent on the efficiency of each round (5). Only very minor changes in the efficiency of each round of amplification have profound effects on the overall yield after many rounds. For example, if the efficiency of the reaction with the perfectly matched primer is 90% and with the mismatched primer is 60%, there would be 73-fold more product produced in the reaction with perfectly matched primer than with the mismatched primer.

The ASPCR should find application in the fields of genetic diagnosis, carrier screening, HLA typing, and any other nucleic acid-based diagnostic in which the precise DNA sequence of the priming site is diagnostic for the target. In the case of HLA typing, recent advances have used PCR amplification followed by allele-specific oligonucleotide hybridization for the determination of *DR*, *DQ*, and *DP* alleles (6, 20–22). It should be possible to use ASPCR for the direct analysis of HLA types.

We have recently proposed a process for the simultaneous determination of multiple polymorphic loci based on the concept of producing locus-specific amplification products each with a unique length (23). In such a system, since ASPCR would produce allele-specific products, the simultaneous analysis of the genotype of the target DNA at multiple loci should be possible.

This work was supported by Grant DCB-8515365 from the National Science Foundation (R.B.W.). D.Y.W. is a M.D./Ph.D. candidate at Loma Linda University. R.B.W. is a member of the Cancer Center of the City of Hope (NIH CA33572). L.U. is a fellow of AIRC (Associazione Italiana per la Ricerca sul Cancro).

1. Kan, Y. W. & Dozy, A. M. (1978) *Lancet* ii, 910–912.
2. Geever, R. F., Wilson, L. B., Nallaseth, F. S., Milner, P. F., Bittner, M. & Wilson, J. T. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5081–5085.
3. Chang, J. C. & Kan, Y. W. (1982) *N. Engl. J. Med.* **307**, 30–32.
4. Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L. & Wallace, R. B. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 278–282.
5. Saiki, R. K., Scharf, S., Falcona, F., Mullis, K., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350–1354.
6. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1986) *Nature (London)* **324**, 163–166.
7. Chehab, F. F., Doherty, M., Cai, S., Kan, Y. W., Cooper, S. & Rubin, E. M. (1987) *Nature (London)* **329**, 293–294.
8. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. (1988) *Science* **241**, 1077–1080.
9. Wu, D. Y. & Wallace, R. B. (1989) *Gene*, in press.
10. Wu, D. Y. & Wallace, R. B. (1989) *Genomics*, in press.
11. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
12. Dembek, P., Miyoshi, K. & Itakura, K. (1981) *J. Am. Chem. Soc.* **103**, 706–708.
13. Bell, G. I., Karam, J. H. & Rutter, W. J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5759–5763.
14. Selden, R. F., Howie, K. B., Rowe, M. E., Goodman, H. M. & Moore, D. D. (1986) *Mol. Cell Biol.* **6**, 3173–3179.
15. Yamane, A., Nakagami, S., Kawasoe, T. & Miyoshi, K. (1988) *Nucleic Acids Res.* **20**, 91.
16. Kidd, V. J., Wallace, R. B., Itakura, K. & Woo, S. L. C. (1983) *Nature (London)* **304**, 230–234.
17. Nozari, G., Rahbar, S. & Wallace, R. B. (1986) *Gene* **43**, 23–28.
18. Atkinson, M. R., Deutscher, M. P., Kornberg, A., Russel, A. F. & Moffet, J. G. (1969) *Biochemistry* **8**, 4897–4904.
19. Tindall, K. R. & Kunkel, T. A. (1988) *Biochemistry* **27**, 6008–6013.
20. Morel, P. A., Dorman, J. S., Todd, J. A., McDevitt, H. O. & Trucco, M. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 8111–8115.
21. Angelini, G., Bugawan, T., Delfino, L., Erlich, H. & Ferrara, G. B. (1988) *Hum. Immunol.* **23**, 77.
22. Scharf, S., Saiki, R. & Erlich, H. (1988) *Hum. Immunol.* **23**, 143.
23. Skolnick, M. H. & Wallace, R. B. (1988) *Genomics* **2**, 273–279.

09/623,068

Publication 5)
With RCE Submission
July 18, 2003

Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 8923–8927, November 1990
Genetics

Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay

(DNA amplification/gene detection/genome mapping)

DEBORAH A. NICKERSON*, ROBERT KAISER, STEPHEN LAPPIN, JASON STEWART†, LEROY HOOD, AND ULF LANDEGREN†

Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125

Contributed by Leroy Hood, August 16, 1990

ABSTRACT **DNA diagnostics, the detection of specific DNA sequences, will play an increasingly important role in medicine as the molecular basis of human disease is defined. Here, we demonstrate an automated, nonisotopic strategy for DNA diagnostics using amplification of target DNA segments by the polymerase chain reaction (PCR) and the discrimination of allelic sequence variants by a colorimetric oligonucleotide ligation assay (OLA). We have applied the automated PCR/OLA procedure to diagnosis of common genetic diseases, such as sickle cell anemia and cystic fibrosis (ΔF508 mutation), and to genetic linkage mapping of gene segments in the human T-cell receptor β-chain locus. The automated PCR/OLA strategy provides a rapid system for diagnosis of genetic, malignant, and infectious diseases as well as a powerful approach to genetic linkage mapping of chromosomes and forensic DNA typing.**

The study of DNA sequence variants in humans is playing an important role in diagnosis of genetic and malignant diseases (1, 2). The analysis of DNA polymorphisms also serves as the fundamental tool in attempts to construct genetic linkage maps (3, 4) and in forensic analyses (5, 6). Since the majority of DNA sequence variants and polymorphisms are single nucleotide substitutions (1, 2), diagnostic techniques must accurately discriminate single base changes.

Single base variations in DNA sequences can be detected by a variety of techniques including Southern blot analysis (7) for restriction fragment length polymorphisms, allele-specific oligonucleotide hybridization (8), denaturing gradient gel electrophoresis (9), chemical cleavage of mismatched heteroduplexes (10), conformational changes in single strands (11), and allele-specific priming of the polymerase chain reaction (PCR) (12–14). These techniques have several disadvantages for automating DNA diagnosis, which include the use of radioactivity, the requirement for various hybridization conditions, and the need for electrophoresis or centrifugation.

The analysis of DNA sequence variants has been greatly facilitated by the development of rapid methods to exponentially amplify specific DNA or RNA targets. Diagnostic targets can be amplified by PCR (15–17) or by other available methods (18–21). Amplification generates specific targets with high signal/noise ratios and permits the use of less sensitive nonisotopic reporters in DNA analysis.

An alternative strategy for DNA diagnosis, the oligonucleotide ligation assay (OLA), employs two adjacent oligonucleotides (20-mers), a 5′ biotinylated probe (with its 3′ end at the nucleotide to be assayed) and a 3′ reporter probe (22–24). The two oligonucleotides are hybridized to target DNA and, if there is perfect complementarity, the enzyme DNA ligase covalently joins the 5′ biotinylated probe and the 3′ reporter probe. If the probes and target are mismatched at their junction, a covalent bond is not formed. Capture of the 5′ biotinylated probe on immobilized streptavidin and analysis for covalently linked 3′ reporters determine the nature of the probe–target interaction (matched or mismatched). The ligase assay uses a standard set of conditions to distinguish all nucleotide mismatches, and product analysis does not require electrophoresis or centrifugation (22). In this report, we describe a strategy for automating DNA diagnosis that combines target amplification by PCR with a nonisotopic analysis of DNA sequence variants by OLA.

MATERIALS AND METHODS

Robotic Workstation. A Biomek 1000 workstation (Beckman) equipped with multipipet tools and a multibulk tool was used to perform all pipetting, aspirating, and washing procedures. The workstation has been modified with a solenoid to switch wash solutions during the ELISA. All reagents for sample processing were stored in sterile 96-minitube cassettes.

DNA Samples. DNA from humans with α_1-antitrypsin, β-globin, and cystic fibrosis variants was obtained from F. Heijtmancik (Baylor University), from K. Tanaka (Harbor Hospital) and J. Korenberg (Cedar–Sinai Hospital), and from A. Osher and E. Hsu (Children's Hospital), respectively, and prepared as described (22). DNA for amplification of human T-cell receptor β-chain (TCRβ) gene segments was obtained by gently scraping cells from the lining of the buccal cavity with a sterile toothpick. Buccal cells were dislodged into a minitube containing 10 μl of sterile H_2O, covered with 75 μl of mineral oil, and placed into a 96-minitube cassette for handling by the robotic workstation. Cells were lysed with 20 μl of 0.1 M KOH and 0.1% Triton X-100 at 65°C for 20 min and neutralized with 20 μl of 0.1 M HCl and 0.1% Triton X-100.

Oligonucleotides. Amplification primers and ligation probes were assembled by using standard phosphoramidite chemistry on an Applied Biosystems 380A DNA synthesizer. Ligation probes were modified with a 5′ biotin group as described (15) or chemically phosphorylated with 5′ Phosphate-ON (Clontech) according to the manufacturer's directions. Modified probes were purified by reverse-phase high-performance liquid chromatography. Phosphorylated oligonucleotide probes (500 pmol) were labeled with dUTP-digoxigenin by mixing 100 mM potassium cacodylate, 2 mM $CoCl_2$, 200 μM dithio-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: PCR, polymerase chain reaction; OLA, oligonucleotide ligation assay; TCRβ, T-cell receptor β chain; CFTR, cystic fibrosis transmembrane conductance regulator; V, variable; D, diversity; J, joining; C, constant; STS, sequence-tagged site.

*To whom reprint requests should be addressed.
†Current address: Department of Medical Genetics, University of Upsala, Box 589, Biomedical Center, S-751 23 Upsala, Sweden.

threitol, 2.5 μl of dUTP-digoxigenin (Boehringer Mannheim), and 2 μl of adenosine triphosphate (40 μM) with 70 units of terminal deoxynucleotidyltransferase (Collaborative Research) for 1 hr at 37°C. Free dUTP-digoxigenin was removed by two successive ethanol precipitations.

DNA Amplification. The robotic workstation was programmed to assemble PCR reagents [5 μl containing 20 mM Tris·HCl (pH 8.3), 100 mM KCl, 3 mM $MgCl_2$, 20 ng of bovine serum albumin per ml, the four deoxynucleotide triphosphates each at 400 μM, 0.5 μM amplification primers, 0.1% Triton X-100, and 0.05 unit of *Thermus aquaticus* DNA polymerase per well], genomic DNA (5 μl at 2 ng/μl in sterile distilled H_2O containing 0.1% Triton X-100), and 70 μl of light mineral oil in a flexible U-bottomed 96-well microtiter plate (Falcon). Genomic DNA samples were denatured at 93°C for 4 min and amplified by 40 cycles of 93°C for 30 sec, 55°C [cystic fibrosis transmembrane conductance regulator (CFTR) and TCRα constant (C_α) gene segments] or 61°C (β-globin and α_1-antitrypsin gene segments) for 45 sec, and 72°C for 90 sec in a microtiter plate thermal cycler (MJ Research, Watertown, MA). For amplification of TCRβ gene segments, 15 μl of PCR reagents (as described above) containing all six amplification primers, 15 μl of the lysed buccal samples, and 70 μl of mineral oil were added to a flexible microtiter plate. Targets were denatured at 93°C for 4 min and amplified by 20 cycles of 30 sec at 93°C, 45 sec at 61°C, and 90 sec at 72°C. Five microliters from these reaction mixtures was used to initiate a second round of amplification for each of the individual TCRβ gene segments (40 cycles; 30 sec at 93°C, 45 sec at 61°C, and 90 sec at 72°C).

Ligation Assays. Ligation reaction mixtures were assembled by the robotic workstation. Forty-five microliters of 0.25 M NaOH containing 0.1% Triton X-100 was added to amplified DNA samples. Ligation probes (200 fmol each) in 10 μl of 2× ligase buffer (100 mM Tris·HCl, pH 7.5/20 mM $MgCl_2$/2 mM spermidine/2 mM adenosine triphosphate/10 mM dithiothreitol) and 50% formamide were added to a U-bottomed 96-well microtiter plate. DNA samples were neutralized with 45 μl of 0.25 M HCl and six 10-μl aliquots were added to the microtiter plate containing the ligation probes. Samples were covered with 70 μl of mineral oil, denatured at 93°C for 2 min, cooled, and returned to the workstation for the addition of 5 μl of T4 DNA ligase (5 units/ml) (Amersham) in 1× ligase buffer. Ligations were done at room temperature (RT) for 15 min. Reactions were stopped by adding 10 μl of 0.25 M NaOH per well and, after 2 min at RT, 4 μl of 3 M sodium acetate (pH 6.5) per well. Samples were transferred to a 96-well flat-bottomed microtiter plate (Falcon) coated with streptavidin [60 μl of streptavidin (100 μg/ml) or avidin (100 μg/ml) (Vector Laboratories) for 1 hr at 37°C] and blocked 20 min (RT) before use with 200 μl of 100 mM Tris·HCl, pH 7.5/150 mM NaCl/0.05% Tween 20 (buffer A) per well with 0.5% dry milk and 100 μg of salmon sperm DNA per ml. Biotinylated probes were captured at RT for 30 min, and the plate was washed twice with 0.01 M NaOH and 0.05% Tween 20 and once with buffer A. Thirty microliters of anti-digoxigenin antibodies (diluted 1:1000; Boehringer Mannheim) in buffer A with 0.5% dry milk was added to each microtiter well. Plates were incubated 30 min (RT) and washed six times with buffer A. Substrate (30 μl of BRL ELISA amplification system per well) was added, the plates were incubated 15 min (RT), and 30 μl of amplifier was added. Spectrophotometric absorbances were taken at 490 nm by a Bio-Tek (Burlington, VT) plate reader and absorbances were directly entered into an IBM-XT computer.

Linkage Analysis. Observed haplotype frequencies were calculated for genetic linkage analysis of TCRβ gene segments with a myriad haplotype program (25). The probability of linkage disequilibrium was calculated based on the χ^2 distribution of the Q statistic described by Hedrick *et al.* (26).

RESULTS

The Automated PCR/OLA Strategy. Our strategy for automated gene analysis is shown in Fig. 1. A Biomek 1000 robotic workstation was used to (*i*) prepare targets and assemble reagents for DNA amplification, (*ii*) mix and ligate 5′ biotinylated probes and 3′ digoxigenin-labeled reporter probes on amplified DNA targets using T4 DNA ligase, (*iii*) capture 5′ biotinylated probes on streptavidin-coated microtiter plates, (*iv*) wash plates, and (*v*) detect the digoxigenin reporter coupled to biotin-labeled probes by an ELISA. Altogether, processing time for 96 samples from entry to computer read-out takes <7 hr. Overnight amplification permits processing of ligation assays from 192 DNA samples in a single day (1200 reactions, triplicates for two alleles).

Amplification Primers and Ligation Probes. A panel of amplification primers and ligation probes for known sequence variants in human DNA have been synthesized (Table 1). Two sets of probes detect mutations that cause common genetic diseases in homozygous individuals, sickle cell anemia and CF (27, 28). Another set detects a common mutation in the α_1-antitrypsin gene that, in homozygous individuals, leads to a predisposition for cirrhosis of the liver in childhood and emphysema in adults (29). The remaining probes detect

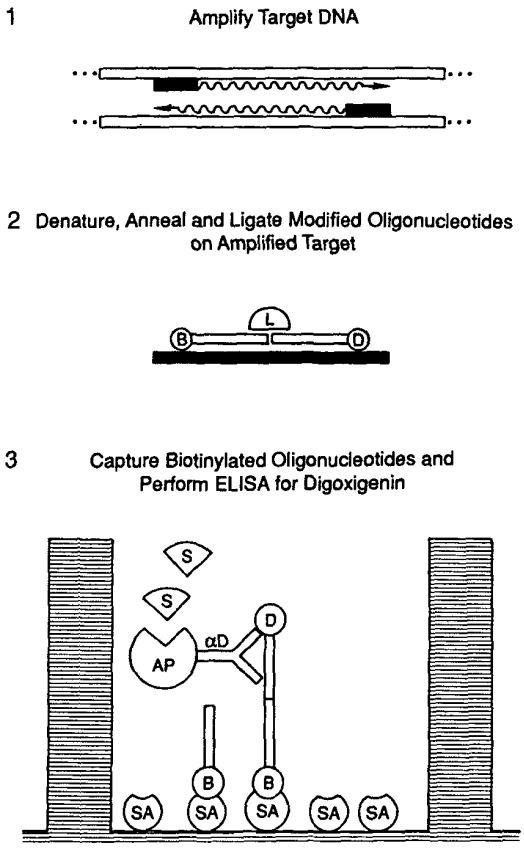


FIG. 1. Schematic diagram of the steps in the automated PCR/OLA procedure performed with a robotic workstation. The assay contains three steps: 1, DNA target amplification; 2, analysis of target nucleotide sequences with biotin (B)-labeled and digoxigenin (D)-labeled oligonucleotide probes and T4 DNA ligase (L); 3, capture of the biotin (B)-labeled probes on streptavidin (SA)-coated microtiter wells and analysis for covalently linked digoxigenin (D) by using an ELISA procedure with alkaline phosphatase (AP)-conjugated anti-digoxigenin (αD) antibodies and a substrate (S).

Table 1. Nucleotide sequence of the amplification primers and ligation probes used in automated DNA analysis

Genomic region amplified	Amplification primers	Ligation probes		Target detected by ligation probes
		Biotin-labeled probe	Reporter-labeled probe	
β-Globin	CAACTTCATCCACGTTCACCTTGCC	1. B-ATGGTGCACCTGACTCCTGA	pGGAGAAGTCTGCCGTTACTG-D	1. βA
	AGGGCAGGAGCCAGGGCTGGG	2. B-ATGGTGCACCTGACTCCTGT		2. βS
α_1-Antitrypsin	TCAGCCTTACAACGTGTCTCTGCTT	1. B-GGCTGTGCTGACCATCGACG	pAGAAAGGGACTGAAGCTGCT-D	1. M
	GTATGGCCTCTAAAAACATGGCCCC	2. B-GGCTGTGCTGACCATCGACA		2. Z
CFTR	CAGTGGAAGAATGGCATTCTGTT	1. B-ATTAAAGAAAATATCATCTT	pTGGTGTTTCCTATGATGAAT-D	1. Non-F508
	GGCATGCTTTGATGACGCTTCTG	2. B-ACCATTAAAGAAAATATCAT		2. ΔF508
Cα	CCTTGAAGCTGGGAGTGG	1. B-GAAACGAAGAAACTGAGGCCA	pCACAGCTAATGAGTGAGGAAGA-D	1. C_α3A
	GAGCTAAGAGAGCCGTACTGG	2. B-GAAACGAAGAAACTGAGGCCC		2. C_α3B
V_β6.71	AAGGGAAAGGATGTAGAG	1. B-TTTACTGGTACCGACAGAGC	pCTGGGGCAGGGCCTGGAGTT-D	1. V_β6.71A
	CTGGCACAGAGATACACGGCC	2. B-TTTACTGGTACCGACAGAGG		2. V_β6.71B
V_β6.72	AAGGGAAAGGATGTAGAG	1. B-TCTGCAGAGAGGACTGGGGG	pATCCGTCTCCACTCTGACGA-D	1. V_β6.72A
	CTGGCACAGAGATACACGGCC	2. B-TCTGCAGAGAGGACTGGGGA		2. V_β6.72B
V_β1	GAGTCACACAAACCCCAAAGCACCT	1. B-AGGCCTCCAGTTCCTCATTCAG	pTATTATAATGGAGAAGAGAGAGCA-D	1. V_β1A
	GCTGCTGGCACAGAAATACAAAGCT	2. B-AGGCCTCCAGTTCCTCATTCAC		2. V_β1B
C_β	CATTATGGTCCTTTCCCGG	1. B-ACCAGGACCAGACAGCTCTC	pAGAGCAACCCTAGCCCCATTAC-D	1. C_β3A
	AGCTCCACGTGGTCGGGGT	2. B-ACCAGGACCAGACAGCTCTT		2. C_β3B

Ligation reactions were performed with a mixture of a biotin-labeled and reporter-labeled probe for each specific allele.

polymorphisms in the human TCRβ and TCRα loci (refs. 30 and 31; C. Whitehurst, P. Charmley, L.H., and D.A.N., unpublished data). Most of these probes detect single nucleotide substitutions in a specific DNA target. However, one set of probes detects a 3-base-pair (bp) deletion in the gene encoding CFTR (28) and represents a model for the detection of sequence deletions by OLA.

Analysis of DNA Sequence Variants. As a model for DNA diagnosis by the PCR/OLA procedure, we obtained genomic DNAs from 32 individuals of known genotype. The robotic workstation was used to assemble PCR reagents and genomic DNA samples in a 96-well microtiter plate. After amplification, ligations were performed in triplicate for each allele, and the immobilized probes were analyzed for the presence of digoxigenin. An example of a microtiter plate obtained from this process is shown in Fig. 2. Amplified targets from homozygous and heterozygous individuals for the indicated nucleotide substitutions (β-globin, α_1-antitrypsin, and TCR C_α) or deletion (CFTR) were used. The assay clearly identifies which alleles 1 and/or 2 (Table 1) were present in each of the amplified samples (Fig. 2). Fig. 3 shows the mean absorbances obtained from ligation assays on amplified DNA targets from eight different individuals for each of the analyzed gene segments (32 individuals altogether). Mean absorbances from different individuals ranged from 0.38 to 1.17. We have found that mean absorbances from the ligation assays reflect the amount of target present in an amplified DNA sample. In this regard, the colorimetric assay is quite sensitive and can detect 3 fmol of ligated product (data not shown). The high signal/noise ratios (10:1–200:1) obtained with this procedure also permit simple data processing to define the genotype of an amplified DNA sample by calculating the ratio of the mean absorbance for each allele in the ligation assay. Furthermore, since the outcome of the PCR/OLA procedure is based on the mean absorbance of triplicate ligation reactions, the chance of error arising from spurious false-negative or false-positive wells is also minimized (false-negative or false-positive wells < 0.2% in 4000 reactions; data not shown).

Genetic Linkage Analysis of TCRβ Genes. The automated PCR/OLA protocol has been extended to include the preparation of DNA samples by the robotic workstation. Amplified DNA targets from human buccal samples were used to determine the frequency and genetic linkage of four DNA sequence polymorphisms in the human TCRβ locus as shown in Fig. 4. The human TCRβ locus is composed of several gene segments, variable (V), diversity (D), end joining (J), and constant (C) genes, which span >600 kilobases (kb) of DNA (Fig. 4) (32, 33). Using data obtained from the automated PCR/OLA procedure on these 96 samples, we found that two V_β6.7 polymorphisms were in complete linkage disequilibrium ($P < 10^{-14}$). This finding was not surprising since these variants are separated by a small physical distance (100 bp). Although the exact location of the V_β6.7 gene segment in the TCRβ locus is not known, analysis of available cosmid and YAC clones by gene-specific PCR suggests that V_β6.7 is probably located 5′ to the V_β1 gene segment. The three TCR polymorphisms (V_β6.7, V_β1, and C_β), physically spanning at least 600 kb, appeared to be in linkage equilibrium with one another. Indeed, the expected haplotype frequencies calculated assuming linkage equilibrium were very close to those observed ($P < 0.81$) (Table 2). These findings confirm those recently reported in a study of TCR polymorphisms detected as restriction fragment length polymorphisms and may sug-

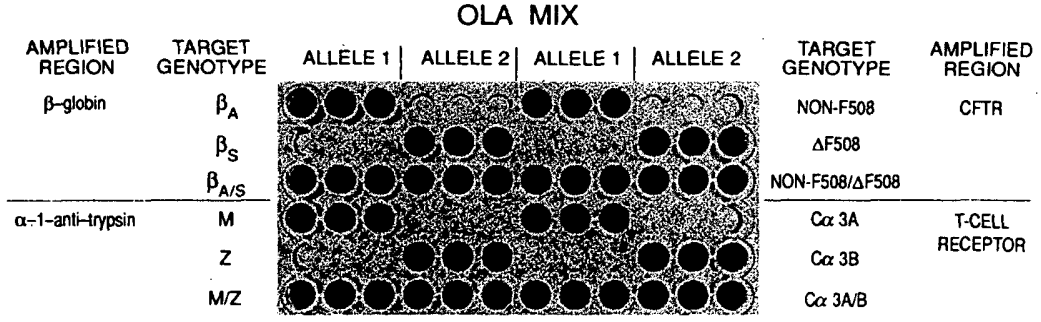


FIG. 2. Amplified DNA targets obtained from genomic DNA samples were analyzed in triplicate by using the indicated combinations of ligation probes (alleles 1 and 2 as described in Table 1) for each specified gene segment. Wells containing digoxigenin form a magenta-colored product and indicate complementarity between the ligation probes and amplified DNA target.

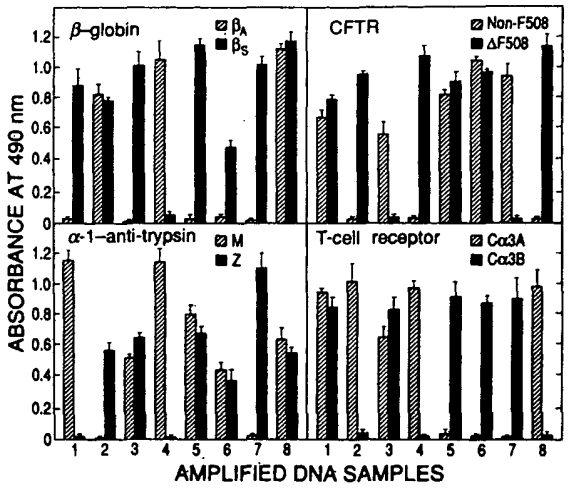


FIG. 3. Mean spectrophotometric absorbances (+1 SD) from triplicate ligation reactions performed by the automated PCR/OLA procedure on amplified DNA samples obtained from eight donors for each gene analyzed (32 DNA samples total).

gest that hot spots of recombination exist in the TCRβ locus (34).

DISCUSSION

Automated analysis of DNA polymorphisms and variants by PCR/OLA has many advantages over existing approaches to DNA diagnostics. Small numbers of cells (cheek scraping) or DNA samples (10 ng) are sufficient for analysis. Only small fragments of DNA (a few hundred base pairs) are required. Therefore, partially degraded DNA is still useful. The reagents are stable and easily obtained, and nonisotopic reporter groups are used. The entire assay is performed in microtiter wells, thus avoiding the use of centrifugation or electrophoresis. The assay yields high signal/noise ratios and a simple readout that is easily transferred to a computer for storage and analysis; no measurements of DNA fragment sizes are necessary. All of the tested sequence variants (nucleotide transitions and transversions, and a deletion) could be discriminated by OLA using a standard set of conditions. The initial PCR amplification facilitates the discrimination of polymorphisms in individual members of a multigene family (e.g., the TCR $V_\beta 6.7$ gene segment is one of nine highly similar members of the $V_\beta 6$ subfamily). The two successive levels of sequence discrimination, PCR and then OLA, enhance signal/noise ratios and reduce the likelihood of error, particularly in the analysis of polymorphisms in multigene families. The steps in the assay are automatable, eliminating the need for human intervention (and possible mistakes) in a tedious and repetitious process. With automation, high throughput is possible. At present, we can process 1200 ligation reactions per day with a single operator and robotic workstation, and, in the near future, further automation with a robotic arm will permit processing of 6000 reactions per day.

Table 2. TCR haplotypes

Haplotype $V_\beta 6.7$ $V_\beta 1$ C_β	Observed	Expected
AAA	64	69
AAB	40	36
ABA	11	14
ABB	9	7
BAA	40	36
BAB	17	19
BBA	6	7
BBB	5	4

Expected haplotypes were calculated assuming random allelic association—e.g., AAA = 0.66 × 0.83 × 0.66 × 192 = 69.

The automated PCR/OLA assay can be applied in many different basic research and clinical areas. Genetic diseases fall into several different categories including the common and widespread mutations of sickle cell disease, α_1-antitrypsin or CF, and newly arising spontaneous mutations such as Lesch–Nyhan disease (35). Clearly, PCR/OLA facilitates the analysis of the common mutations, either in screening at-risk members of families with diseases or for more general carrier screening purposes. Rapid techniques are being developed to identify the sequence variations of newly arising mutations (35, 36). Once identified, the combined PCR/OLA procedure can be used to follow the inheritance of these specific mutations in affected families. Many genes cause a predisposition toward disease. This is true of the α_1-antitrypsin mutation described above. Recently, it has been demonstrated that certain TCR and HLA haplotypes may predispose humans to certain autoimmune diseases such as multiple sclerosis (37–39). Therapeutic strategies are being developed to circumvent these predispositions (40–42). Therefore, automated screening may be useful in the near future to identify the genes associated with disease predis-

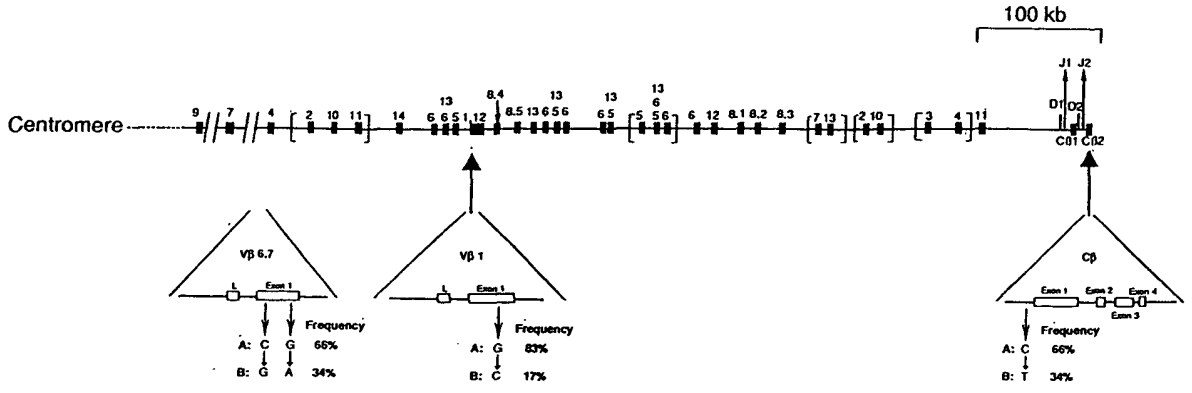


FIG. 4. Schematic diagram of the human TCRβ locus giving the relative order of the V, D, J, and C gene segments. DNA polymorphisms in three indicated gene segments were analyzed in 96 individuals. Their location, where known, is shown (arrow up). The nucleotide substitutions analyzed and the frequency for each variant in these samples are shown.

positions in which some form of preventive therapy can be initiated.

The automated PCR/OLA procedure provides a powerful approach to high-resolution genetic linkage mapping of the human genome or other complex genomes. For this approach, sequence-tagged sites (STSs) (43) from specific chromosomal regions (e.g., the TCRβ locus) or from a specific chromosome (e.g., STSs obtained from random clones of a flow-sorted chromosome library) would be scanned for internal DNA sequence polymorphisms (9–11) to obtain a set of polymorphic STSs. Once acquired, polymorphic STSs can be rapidly ordered by analysis of large multigeneration families or by single-sperm typing (44, 45) using the automated PCR/OLA system.

The availability of human polymorphic STSs will also provide a set of markers for automated forensic typing. For example, with a set of maximally informative biallelic markers (50:50 distribution in random mating populations) from each of the 22 human autosomes, the probability that two individuals would have identical DNA fingerprints—i.e., the same set of the 44 alleles—is ≈ 1 in 10^{10}. The automated PCR/OLA procedure eliminates most of the limitations associated with forensic typing by conventional Southern blot analysis (e.g., the measurement of DNA fragment sizes, the requirement for high quality DNA, and the use of radioisotopes).

Other applications for automated DNA diagnosis by the PCR/OLA procedure include HLA typing, the analysis of recessive or dominant oncogenes, and the identification of infectious pathogens. The use of commercially available thermostable ligases and automated ligation amplification reactions in the direct detection of single copy genes can also be explored. Moreover, multiple nonisotopic reporter groups may be developed that will be simultaneously analyzed in a single microtiter well. This raises the possibility of multiplexing the OLA procedure to the point where initially both alleles can be analyzed together and eventually multiple biallelic loci can be typed in a single well. These and other improvements, such as a single instrument to perform the entire analysis, will greatly increase the throughput and potential applications of automated DNA diagnostics.

We thank Ms. Anna Marie Aquinaldo and Drs. Conrad Sevilla and Suzanna Horvath for their assistance in oligonucleotide synthesis; all our cheek scraping donors; and Drs. P. Charmley, C. Delahunty, T. Hunkapiller, B. Koop, M. Nishimura, L. Rowen, and D. Zaller for their careful review of the manuscript. This work was supported by the Whittier Foundation, National Science Foundation Grant DIR 8809710, and National Institutes of Health Grant HG 00084.

1. Landegren, U., Kaiser, R., Caskey, C. T. & Hood, L. (1988) *Science* **242**, 229–237.
2. Antonarakis, S. E. (1989) *N. Eng. J. Med.* **320**, 153–163.
3. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32**, 314–331.
4. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephen, K., Keith, T. P., Bowden, D. W., Smith, D. R., Lander, E. S., Botstein, D., Akots, G., Rediker, K. S., Gravius, T., Brown, V. A., Rising, M. B., Parker, C., Powers, J. A., Watt, D. E., Kauffman, E. R., Brecker, A., Phipps, P., Muller-Kahle, H., Fulton, T. R., Ng, S., Schummn, J. W., Braman, J. C., Knowlton, R. G., Barker, D. F., Crooks, S. M., Lincoln, S. E., Daly, M. J. & Abrahamson, J. (1987) *Cell* **51**, 319–337.
5. Lander, E. S. (1989) *Nature (London)* **339**, 501–505.
6. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, H., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & White, R. (1987) *Science* **235**, 1616–1622.
7. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
8. Conner, B. J., Reyes, A. A., Morin, C., Itakura, K., Teplitz, R. L. & Wallace, R. B. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 278–282.
9. Meyers, R. M., Maniatis, T. & Lerman, L. S. (1987) *Methods Enzymol.* **155**, 501–527.
10. Cotton, R. G. H., Rodriques, N. R. & Campbell, R. D. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4397–4401.
11. Orita, M., Iwahana, H., Kanazawa, H., Hayashi, H. & Sekiya, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2766–2770.
12. Chehab, F. F. & Kan, Y. W. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9178–9182.
13. Newton, C. R., Graham, A., Heptinstall, L. C., Powell, S. J., Summers, C., Kalsheker, N., Smith, J. C. & Markham, A. F. (1989) *Nucleic Acids Res.* **17**, 2503–2516.
14. Wu, D. Y., Ugozzoli, L., Pal, B. K. & Wallace, R. B. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2757–2760.
15. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985) *Science* **230**, 1350–1354.
16. Mullis, K. B. & Faloona, F. (1987) *Methods Enzymol.* **155**, 335–350.
17. Saiki, R. K., Gelfand, D., Stoffel, S., Scharf, S., Higuchi, R., Horn, G., Mullis, K. & Erlich, H. (1988) *Science* **239**, 487–491.
18. Wu, D. Y. & Wallace, R. B. (1989) *Genomics* **4**, 560–569.
19. Lizardi, P., Guerra, E., Lameli, H., Luna-Tussie, T. & Kramer, F. R. (1988) *BioTechnology* **6**, 1197–1202.
20. Guatelli, J. C., Whitefield, K. M., Kwoh, D. Y., Barringer, K. J., Richman, D. & Gingeras, T. R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1874–1878.
21. Kwok, D. Y., Davis, G. R., Whitefield, K. M., Chappelle, H. L., DiMichele, L. J. & Gingeras, T. R. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 1173–1177.
22. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. (1988) *Science* **241**, 1077–1080.
23. Alves, A. M. & Carr, F. J. (1988) *Nucleic Acids Res.* **16**, 8723.
24. Wu, D. Y. & Wallace, R. B. (1989) *Gene* **76**, 245–254.
25. MacLean, C. J. & Morton, N. E. (1985) *Genet. Epidemiol.* **2**, 263–272.
26. Hedrick, P. W., Thompson, G. & Klintz, W. (1986) in *Evolutionary Processes and Theory*, eds. Karlin, S. & Nevo, E. (Academic, New York), pp. 583–606.
27. Winslow, R. M. & Anderson, W. F. (1983) in *The Metabolic Basis of Inheritance*, eds. Stanbury, J. B., Wyngaarden, J. B., Frederickson, D. S., Goldstein, J. L. & Brown, M. S. (McGraw-Hill, New York), pp. 1666–1710.
28. Riordan, J. R., Rommens, J. M., Kerem, B.-S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plausic, N., Chou, J.-L., Drumm, M. L., Iannuzzi, M. C., Collins, F. S. & Tsui, L.-C. (1989) *Science* **245**, 1066–1072.
29. Crystal, R. G. (1989) *Trends Genet.* **5**, 411–417.
30. Robinson, M. A. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9422–9426.
31. Li, Y., Szabo, P., Robinson, M. A., Dong, B. & Posnett, D. N. (1990) *J. Exp. Med.* **171**, 221–230.
32. Wilson, R. K., Lai, E., Concannon, P., Barth, R. K. & Hood, L. E. (1988) *Immunol. Rev.* **101**, 149–172.
33. Lai, E., Concannon, P. & Hood, L. (1988) *Nature (London)* **331**, 543–546.
34. Charmley, P., Chao, A., Concannon, P., Hood, L. & Gatti, R. A. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4823–4827.
35. Gibbs, R. A., Nguyen, P. N., McBride, L. J., Koepf, S. M. & Caskey, C. T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 1919–1923.
36. Grompe, M., Muzny, D. M. & Caskey, C. T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 5888–5892.
37. Odum, N., Hyldig-Nielsen, J. T., Morling, N., Sandberg-Wollheim, M., Platz, P. & Svejgarrd, A. (1988) *Tissue Antigens* **31**, 235.
38. Beall, S. S., Concannon, P., Charmley, P., McFarland, H. F., Gatti, R. A., Hood, L. E., McFarlin, D. E. & Biddison, W. E. (1989) *J. Neuroimmunol.* **21**, 59–66.
39. Seboun, E., Robinson, M. A., Doolittle, T. H., Ciulla, T. A., Kindt, T. J. & Hauser, S. L. (1989) *Cell* **57**, 1095–1100.
40. Urban, J. L., Kumar, V., Kono, D. A., Coomey, C., Horvath, S. J., Clayton, J., Ando, D. G., Sercarz, E. E. & Hood, L. E. (1988) *Cell* **54**, 577–592.
41. Archa-Orbea, H., Mitchell, D. J., Timmermann, L., Wraith, D. C., Tousch, G. S., Waldor, M. F., Zamvil, S. S., McDevitt, H. O. & Steinman, L. (1988) *Cell* **54**, 263–273.
42. Zaller, D., Osman, G., Kanagawa, O. & Hood, L. (1990) *J. Exp. Med.* **171**, 1942–1955.
43. Olson, M., Hood, L. E., Cantor, C. & Botstein, D. (1989) *Science* **245**, 1434–1435.
44. Li, H., Gyllensten, U., Cui, X., Saiki, R. K., Erlich, H. A. & Arnheim, N. (1988) *Nature (London)* **335**, 414–417.
45. Cui, X., Li, H., Goradia, T. M., Lange, K., Kazazian, H. H., Galas, D. & Arnheim, N. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9389–9393.

generated a substrate that was extended by the polymerase to a complete 50-bp duplex molecule (Fig. 4). This confirms the result shown in Fig. 2B that Rad1-Rad10 removes the 3′ single-stranded tail, and indicates that Rad1-Rad10 cleavage products contain 3′-OH groups, the required substrate for extension by DNA polymerase. Hence, Rad1-Rad10 endonuclease products are suitable substrates for a necessary subsequent step in both the SSA recombination and NER models.

REFERENCES AND NOTES

1. E. C. Friedberg, W. Siede, A. J. Cooper, in *The Molecular and Cellular Biology of the Yeast* Saccharomyces: *Genome Dynamics, Protein Synthesis and Energetics*, J. Broach, E. Jones, J. Pringle, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), pp. 147–192.
2. A. Aguilera and H. L. Klein, *Genetics* **123**, 683 (1989).
3. H. L. Klein, *ibid.* **120**, 367 (1988).
4. B. J. Thomas and R. Rothstein, *ibid.* **123**, 725 (1989).
5. B. R. Zehfus, A. D. McWilliams, Y. H. Lin, M. F. Hoekstra, R. L. Keil, *ibid.* **126**, 41 (1990).
6. R. H. Schiestl and S. Prakash, *Mol. Cell. Biol.* **8**, 3619 (1988).
7. ———, *ibid.* **10**, 2485 (1990).
8. J. Fishman-Lobell and J. E. Haber, *Science* **258**, 480 (1992).
9. A. M. Bailis, L. Arthur, R. Rothstein, *Mol. Cell. Biol.* **12**, 4988 (1992).
10. M. Biggerstaff, D. E. Szymkowski, R. D. Wood, *EMBO J.* **12**, 3685 (1993).
11. A. J. van Vuuren *et al.*, *ibid.*, p. 3693.
12. M. van Duin *et al.*, *Cell* **44**, 913 (1986).
13. A. J. Bardwell, L. Bardwell, D. K. Johnson, E. C. Friedberg, *Mol. Microbiol.* **8**, 1177 (1993).
14. L. Bardwell, A. J. Cooper, E. C. Friedberg, *Mol. Cell. Biol.* **12**, 3041 (1992).
15. A. E. Tomkinson, A. J. Bardwell, L. Bardwell, N. J. Tappe, E. C. Friedberg, *Nature* **362**, 860 (1993).
16. P. Sung, P. Reynolds, L. Prakash, S. Prakash, *J. Biol. Chem.* **268**, 26391 (1993).
17. A. J. Bardwell, L. Bardwell, A. E. Tomkinson, E. C. Friedberg, data not shown.
18. J.-C. Huang, D. L. Svoboda, J. T. Reardon, A. Sancar, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 3664 (1992).
19. L. Bardwell *et al.*, *ibid.* **91**, 3926 (1994).
20. A. J. Bardwell *et al.*, *Mol. Cell. Biol.* **14**, 3569 (1994).
21. W. J. Feaver *et al.*, *Cell* **75**, 1379 (1993).
22. We have not detected exonuclease activity on single-stranded, double-stranded or 5′-tailed DNA oligonucleotides incubated with Rad1-Rad10, nor on phage DNA (*15*, *28*). However, we cannot formally exclude the possibility that the Rad1-Rad10 activity observed on the 3′-tailed and partial duplex structures in this study is a combined endonuclease-exonuclease with specificity for 3′ tails.
23. Y. Habraken, P. Sung, L. Prakash, S. Prakash, *Nature* **366**, 365 (1993).
24. J. J. Harrington and M. R. Lieber, *Genes Dev.* **8**, 1344 (1994).
25. Recent studies [A. O'Donovan, A. A. Davies, J. G. Moggs, S. C. West, R. D. Wood, *Nature*, in press] have shown that XPG protein (the human homolog of yeast Rad2 protein) is a junction-specific endonuclease that cuts DNA at duplex–5′ single-strand regions.
26. F. Lin, K. Sperle, N. Sternberg, *Mol. Cell. Biol.* **4**, 1020 (1984).
27. B. A. Ozenberger and G. S. Roeder, *ibid.* **11**, 1222 (1991).
28. A. E. Tomkinson, A. J. Bardwell, N. Tappe, W. Ramos, E. C. Friedberg, *Biochemistry* **33**, 5305 (1994).
29. Rad1 and Rad10 proteins were purified as described (*15*, *28*) [L. Bardwell, H. Burtscher, W. A. Weiss, C. M. Nicolet, E. C. Friedberg, *Biochemistry* **29**, 3119 (1990)]. Constituent oligonucleotides for each substrate (see Table 1) were mixed, heated to 95°C, and annealed by cooling to room temperature. Endonuclease reactions were carried out at 37°C for 40 min in 20-μl volumes containing 50 mM tris (pH 8.5), 5 mM $MgCl_2$, 5 mM dithiothreitol, and 1 pmol of substrate DNA. Reactions were stopped, deproteinized, and analyzed by gel electrophoresis and autoradiography. The low specific activity of Rad1-Rad10 enzyme has been previously observed (*15*, *16*).

30. We thank N. Tappe for assistance with protein purification. A.J.B. has previously published under the name A. J. Cooper. Supported by research grants CA12428 from the U.S. Public Health Service (E.C.F.) and grant 3786 from the Council for Tobacco Research (A.E.T.).

12 May 1994; accepted 17 August 1994

Padlock Probes: Circularizing Oligonucleotides for Localized DNA Detection

Mats Nilsson, Helena Malmgren, Martina Samiotaki, Marek Kwiatkowski, Bhanu P. Chowdhary, Ulf Landegren*

Nucleotide sequence information derived from DNA segments of the human and other genomes is accumulating rapidly. However, it frequently proves difficult to use such short DNA segments to identify clones in genomic libraries or fragments in blots of the whole genome or for in situ analysis of chromosomes. Oligonucleotide probes, consisting of two target-complementary segments, connected by a linker sequence, were designed. Upon recognition of the specific nucleic acid molecule the ends of the probes were joined through the action of a ligase, creating circular DNA molecules catenated to the target sequence. These probes thus provide highly specific detection with minimal background.

The application of synthetic oligonucleotides in combination with nucleic acid–specific enzymes has brought simplicity and convenience to molecular genetic analyses. There is, however, a need for methods in which oligonucleotides can be used for localized detection of single-copy gene sequences and for distinction among sequence variants in microscopic specimens. Such methods would help to bridge the analytic gap between specific gene sequences and subcellular structures. We have developed oligonucleotide probe molecules that should be useful for localized detection of specific nucleic acids. These "padlock" probes are composed of two target-complementary segments, connected by a linker that may carry detectable functions. The two ends of the linear oligonucleotide probes are brought in juxtaposition by hybridization to a target sequence. This juxtaposition allows the two probe segments to be covalently joined by the action of a DNA ligase. Because of the helical nature of DNA, circularized probes are wound around the target strand, topologically connecting probes to target molecules through catenation, in a manner similar to padlocks. The requirement for simultaneous hybridization of two different probe segments to target molecules provides for high specificity of detection in complex populations of nucleic acids (*1*). Moreover, the act of ligation permits facile distinction among similar target sequence variants as terminally mismatched probes are poor substrates for ligases (*1*, *2*). Finally, the covalent catenation of probe molecules to target sequences described here results in the formation of a hybrid that resists extreme washing conditions, serving to reduce nonspecific signals in genetic assays.

Probes useful for circularization experiments were constructed by solid phase synthesis of oligonucleotides that contained two hybridizing regions of 20 nucleotides each, connected by a 50-nucleotide-long linker segment (Fig. 1). Phosphate groups were added at the 5′ ends of the molecules as required for enzymatic ligation. Alternatively, residues of hexaethylene glycol (HEG) were incorporated in the linker segment during standard solid phase synthesis (*3*). The HEG residues served to reduce the number of synthetic steps required to span the ends of the two target-complementary segments.

Cyclizable probes were designed to detect a 40-nucleotide target sequence, represented either by an oligonucleotide molecule or by the polylinker sequence of the single-stranded form of the circular cloning vector M13 mp18. Ligation products could be separated by denaturing polyacrylamide gel electrophoresis (Fig. 2A). In the presence of the oligonucleotide target, linear probes were efficiently converted to circular molecules with a

M. Nilsson, H. Malmgren, M. Samiotaki, M. Kwiatkowski, U. Landegren, The Beijer Laboratory, Department of Medical Genetics, Box 589 Biomedical Center, S-75123 Uppsala, Sweden.
B. P. Chowdhary, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Box 7023, S-750 07 Uppsala, Sweden.
*To whom correspondence should be addressed.

distinct rate of migration. Probes interacting with M13 target molecules were converted to a species catenated to and therefore migrating with the large M13 molecule during denaturing gel electrophoresis. As the probes were labeled by the addition of a radioactive phosphate group at the 5′ terminus, only ligated molecules retained their label after treatment with alkaline phosphatase. Circular oligonucleotides are insensitive to digestion with exonuclease VII, which attacks at free 5′ or 3′ ends of DNA strands (4). Depending on how the probes are labeled, phosphatases or exonucleases could be used to remove any signal arising from unreacted probes in various assays, thus reducing background (5).

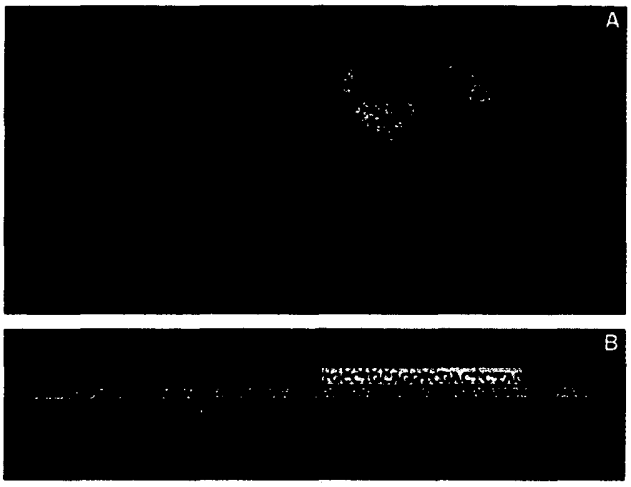


Fig. 1. Structure of a padlock probe interacting with its target sequence. (**A**) Molecular model of the probe-target complex. The molecular model was prepared on a Silicon Graphics workstation, with Insight II (Biosym Technologies). (**B**) Sequence composition of a probe, specific for a segment present in the M13 cloning vector sequence. At the 5′ end of the probe, beginning with a phosphate group, 20 target-complementary nucleotide positions are shown in red. Directly contiguous with these is a linker segment of 50 T residues, shown in green. Finally, the 20 nucleotides at the 3′ end of the probe are yellow. The target sequence is shown in blue.

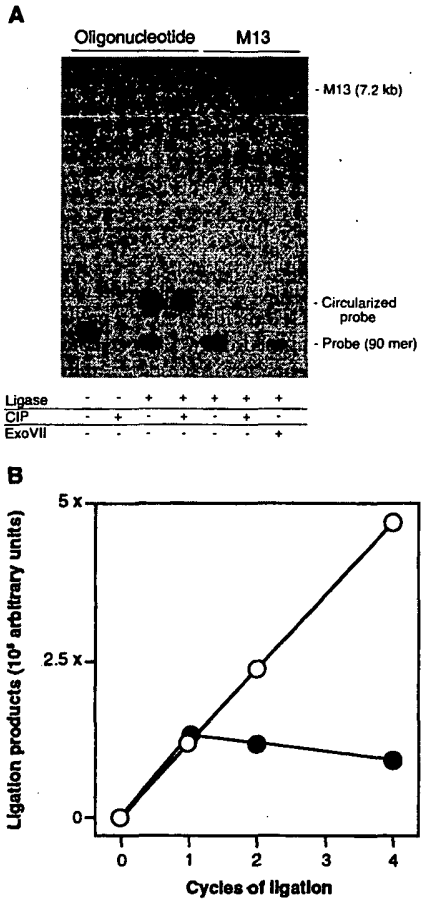


Fig. 2. Analysis by gel electrophoresis of the target-dependent circularization of an oligonucleotide probe. (**A**) A 90-bp oligonucleotide probe (5′ TGCCTGCAGGTCGACTCTAG$(T)_{50}$-CGGCCAGTGCCAAGCTTGCA-3′, see also Fig. 1B) was designed such that its 5′ and 3′ ends would hybridize adjacent to each other to a segment in the polylinker region of the M13 mp18 cloning vector. The probe was gel-purified and 5′-phosphorylated by T4 polynucleotide kinase (New England Biolabs) and γ^{32}P-ATP (3000 Ci/mmol, Dupont). To ensure that most or all 5′ ends were phosphorylated, a second kinase incubation was performed in the presence of a 20-fold excess of adenosine triphosphate (ATP). The labeled probe (6 pmol) was incubated with 3 pmol of either of two different templates: the 7.2-kb, single-stranded, circular M13 mp18 molecule or an oligonucleotide (5′-TTTTTCTAGAGTCGACCTGCAGGCATGCAAGCTTGGCACTGGCCGTTTTT-3′) that contained the same 40-bp target sequence, in 100 μl of 20 mM tris-HCl (pH 8.3), 25 mM KCl, 10 mM $MgCl_2$, 1 mM NAD^+, 0.01% Triton X-100, and 200 U of Ampligase (Epicentre Technologies). The reactions were heated to 90°C (1 min), then cooled to 55°C (5 min) and chilled on ice. Samples (10 μl) were taken from the ligation reactions and treated with either 0.5 U of calf intestinal alkaline phosphatase (CIP; New England Biolabs) or 0.1 U of exonuclease VII (Exo VII; Gibco/BRL). (**B**) The same probe (9 pmol) was subjected to repeated cycles of ligation, separated by heat denaturation steps, in the presence of 0.3-pmol oligonucleotide target (open circles) or the circular single-stranded target molecule (filled circles). Radioactive ligation products, accumulated after the indicated number of cycles, were separated by gel electrophoresis on a 6% denaturing polyacrylamide gel and quantitated with a Phosphorimager (Molecular Dynamics).

We also investigated the consequences of cyclically repeating the probe hybridization and ligation reaction. The amount of cyclized probe molecules increased linearly with the number of ligation cycles when a short oligonucleotide target was used (Fig. 2B). By contrast, under the same conditions the maximal number of probes were bound to the closed, circular M13 target molecule in a single ligation cycle; thereafter the signal decreased, probably because of scission of the single-stranded target molecule during heat denaturation. Thus, a single probe may be catenated to each circular target molecule. This indicates that circularized probe molecules, constrained to one-dimensional diffusion along the target strand during heat denaturation, rapidly occupy the correct target sequence before new probes bind to this sequence when the temperature is lowered. Repeated cycles of ligation can, however, increase the probability that any target sequence will be detected by probe molecules specific for that target, particularly when allele-specific probes are used to distinguish among sequence variants.

Investigators can use oligonucleotide probe ligation reactions to distinguish among related DNA sequences by studying their ability to serve as templates for ligation of oligonucleotides complementary to one or the other sequence variant (1). Whereas probes specific for one of the two sequence variants may hybridize stably to either of the two sequences, only target molecules correctly base-paired to the juxtaposed ends of the probes can assist in the ligation. We investigated the capacity of the padlock probes to distinguish between a normal and a mutant DNA sequence in plasmid clones immobilized on nylon membranes (Fig. 3). Plasmids containing the ΔF508 variant of the cystic fibrosis transmembrane conductance regulator (CFTR) gene or the corresponding normal gene segment were spotted on nylon membranes and subjected to probe hybridization and ligation. The mutation removes 3 base pairs (bp) (6) corresponding to the 3′ end of the circularizable probe. Probe molecules specific for the normal sequence gave rise to a signal only when reacted with the normal sequence but not with the ΔF508 variant of the CFTR gene when probe ligation was followed by denaturing washes in 0.2 M NaOH for 5 min. This stringent wash (to interrupt hybridization between DNA molecules) permitted effi-

cient distinction between the allelic variants, as only cyclized probes remain bound to the membrane. By contrast, a stringent but nondenaturing wash of the same probes in a solution of 2% SDS in 0.1 × standard saline citrate (SSC) gave poor distinction between the two target sequences. Because signal strength is preserved under conditions that prevent hybridization between complementary DNA strands, nonspecifically trapped probe molecules may be efficiently removed, resulting in a reduction of the level of background in gene detection reactions.

As indicated in Fig. 2B, circularized probe molecules are free to travel considerable distances along the target strands during denaturing washes. To measure the distance traveled, probe-cyclization reactions were carried out on equivalent numbers of covalently closed target molecules or molecules that had been linearized at variable distances from the probe-complementary sequence before being immobilized on nylon membranes (Fig. 3B). Few probe molecules that were cyclized around target strands interrupted approximately 150 nucleotides from the probe-complementary sequence remained after denaturing washes. By contrast, strands digested 850 nucleotides from the probe-complement retained similar numbers of probes as did uninterrupted strands. The greater preservation of signal upon denaturing washes of probes bound to the longer linear target molecules probably reflects the increased likelihood that target molecules were cross-linked to the membrane on both sides of the site where the probe was catenated. This trapping of circularized probes by catenation to linear target molecules, in combination with the specific detection afforded by the requirement that two different probe segments simultaneously react with the target sequence, should be of value in procedures such as DNA blotting or for screening genomic libraries with short probe sequences.

Currently, oligonucleotide probes find limited applications for in situ analysis of gene sequences in metaphase chromosomes. This is a consequence of problems both with specificity of detection and sensitivity of visualization. A circularizable probe, specific for a repeated centromeric motif characteristic of human chromosome 12 (*7*), was used for in situ hybridization followed by ligation in human metaphase chromosome preparations. A wide range of washing conditions, including ones that remove specifically hybridizing oligonucleotide or longer probes preserved signals from in situ circularized probe molecules and permitted efficient distinction from alphoid repeat sequences present on other human chromosomes (Fig. 4). Given sufficiently sensitive techniques for detection of probe molecules, the high specificity of padlock probes in conjunction with the reduced nonspecific background observed should permit detection of short, single-copy DNA sequences in human chromosomes. Increased signal could be obtained by secondary ligation of detectable molecules to the linker segment of bound probes. Thus, oligonucleotide probes could be used to screen for the presence of known mutations in loci distributed along the chromosomes, by means of color-coded probes specific for normal and mutant sequence

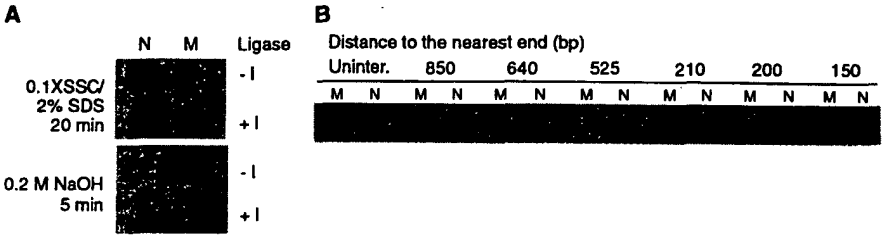


Fig. 3. Distinction of target DNA molecules immobilized on nylon membranes by means of a circularizable probe. (**A**) Fifteen femtomoles of two plasmids containing the normal or a 3-bp deleted variant of the CFTR gene were spotted on nylon membranes (PALL). The filters were treated with 0.1% SDS in boiling water and left for 10 min at room temperature; filters were then washed twice with phosphate-buffered saline (PBS) (*9*) to remove plasmids that had not been fixed to the membrane. Thirty femtomoles of a circularizable probe (5′-P TGGTGTTTCCTATGA($(HEG)_2$C-B)$_4$$(HEG)_2$AAGAAATATCATCTT-3′) per microliter was hybridized to the membranes for 30 min in 5 × SSPE (*9*); 5 × Denhardt's solution (*9*), and salmon sperm DNA (500 μg/ml). The probe contained NH_2-modified C residues to which biotin had been coupled by means of a biotin-NHS ester (Clonetech Laboratories) as described (*10*). Next, the membranes were incubated for 1 hour at room temperature in a solution of 10 mM tris, pH 7.5, 10 mM $Mg(Ac)_2$, 50 mM KAc, 0.2 M NaCl, 1 mM ATP, and 0.15 U of T4 DNA ligase per microliter (Pharmacia). The membranes were washed in a solution of 2% SDS in 1 × SSPE for 30 min, next in either 2% SDS in 0.1 × SSC for 30 min for a stringent wash, or, for a denaturing wash, in 0.2 M NaOH for 5 min, and then in 1 × SSPE, 2% SDS, for 30 min. A signal was generated by incubating the membranes for 5 min in streptavidin–horseradish peroxidase conjugate (0.05 μg/ml; Boehringer Mannheim) in 2 × SSPE, 2% SDS, rinsing in PBS for 30 to 60 min, and then soaking in ECL solution (Amersham) for 1 min. The chemoluminescent signal was recorded on X-omat-S film. (**B**) Plasmids containing the normal (N) or mutant (M) variants of the target molecules were digested with restriction enzymes at the indicated distances from the sequence complementary to the probe or were left undigested. After immobilization on nylon membranes, the plasmids were probed by hybridization with a circularizable oligonucleotide, followed by a ligation step and a denaturing wash in 0.2 M NaOH.

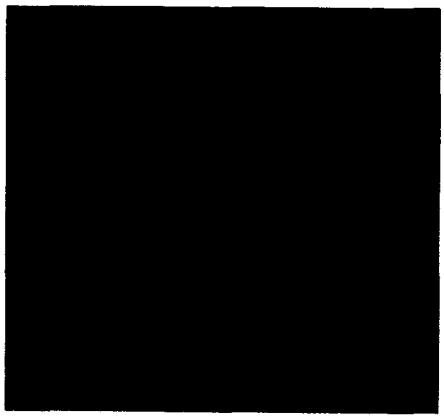


Fig. 4. Detection of a chromosome 12–specific repeated sequence in human metaphase chromosomes, by in situ hybridization and ligation of a biotinylated circularizable probe. Metaphase chromosome preparations were obtained from a human lymphocyte culture by standard techniques of colcemide treatment, hypotonic shock, and fixation in methanol + acetic acid. In situ hybridization and ligation were performed by a modification of the procedure described (*11*). The slides were treated with ribonuclease A at 200 μg/ml in 2 × SSC (*9*) for 1 hour at 37°C, dehydrated in a series of 70, 90, 95, and 99% ice-cold ethanol washes for 2 min each, and air-dried. The chromosome preparations were then denatured in 70% formamide, 2 × SSC at 70°C for 2 min; immediately dehydrated in a series of 70, 90, 95, and 99% ice-cold ethanol washes for 2 min each; and air-dried. Circularizable probe (10 fmol/μl) specific for an alphoid repeat–motif present on chromosome 12 (5′-P AAATCTCCAACTGGAAACTG ($(HEG)_2$(C-B))$_7$$(HEG)_2$ ATTTGGTCTCAAAGTGATTG-3′) was hybridized for 18 hours at 37°C 2 × SSC, 20% formamide and salmon sperm DNA (1 μg/μl) in a 25-μl volume on each slide. A 5-min wash in 2 × SSC at 37°C and a brief wash in 10 mM tris, pH 7.5, 10 mM $Mg(Ac)_2$, 50 mM KAc, 10 mM ATP preceded ligation in the same buffer, containing T4 DNA ligase (0.085 U/μl) for 1 hour at 37°C. The slides were washed twice in 2 × SSC with 20% formamide at 37°C for 5 min each, followed by two washes in 2 × SSC and once in PN buffer (0.1 M NaH_2PO_4 0.1% NP-40, adjusted to pH 8.0 with 0.1 M Na_2HPO_4) at 37°C, 5 min each. Bound probes were visualized by means of fluorescein-labeled avidin, followed by a layer of biotinylated antibodies against avidin, both at 5 μg/ml (Vector Laboratories), and a second layer of fluoresceinated avidin. All incubations were performed in PN buffer containing 5% nonfat milk at 37°C for 20 min followed by three washes in PN buffer at room temperature for 5 min each. The metaphase chromosomes were stained with propidium iodide and photographed with a Nikon Axiofot microscope.

variants. Furthermore, probe cyclization reactions depend on an intramolecular reaction as opposed to reaction between pairs of independent probe molecules as in amplification by the polymerase chain reaction. Thus, there should be fewer problems with nonspecific reactions resulting from interactions between noncognate pairs of probe segments with cyclizable probes. The present probe design should permit the simultaneous analysis of multiple gene sequences in a DNA sample.

In conclusion, the nucleic acid probe presented here permits highly specific detection of nucleotide sequences and, although the target is not amplified, highly sensitive detection is possible through efficient reduction of nonspecific signal. Circularizable probes should be applicable in a number of other contexts, including the detection of specific RNA molecules expressed in tissue sections as T4 DNA ligase can assist in ligation reactions involving RNA strands (*8*). Moreover, immobilized padlock probes could be useful for preparative purposes, such as trapping circular target molecules from solution when screening gene libraries.

REFERENCES AND NOTES

1. U. Landegren, R. Kaiser, J. Sanders, L. Hood, *Science* **241**, 1077 (1988); A. M. Alves and F. J. Carr, *Nucleic Acids Res.* **16**, 8723 (1988); F. Barany, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 189 (1991).
2. D. Y. Wu and R. B. Wallace, *Gene* **76**, 245 (1989).
3. A. Jäschke, J. P. Fürste, D. Cech, V. A. Erdmann, *Tetrahedron Lett.* **34**, 301 (1993).
4. G. Prakash and E. T. Kool, *J. Am. Chem. Soc.* **114**, 3523 (1992); N. G. Dolinnaya *et al.*, *Nucleic Acids Res.* **21**, 5403 (1993).
5. The upper faint bands observed in lanes 3 and 4 probably represent small amounts of linear dimer molecules, appearing as a consequence of ligation of one end each of two different probe molecules. This material proved susceptible to exonuclease, digestion. The extra lower bands in these lanes were not reproducible between experiments. Small amounts of uncatenated, circular probes appearing in lane 7 most likely were a consequence of endonuclease activity in the exonuclease preparation. With increasing amounts of exonuclease, catenated probes are lost and more free circular probes appear (M. Nilsson *et al.*, unpublished data).
6. J. R. Riordan *et al.*, *Science* **245**, 1066 (1989).
7. H. F. Willard and J. S. Waye, *Trends Genet.* **3**, 192 (1987). A. G. Matera and D. C. Ward, *Hum. Mol. Genet.* **7**, 535 (1992). A. Baldini *et al.*, *Am. J. Hum. Genet.* **46**, 784 (1990).
8. N. P. Higgins and N. R. Cozzarelli, *Methods Enzymol.* **68**, 50 (1979).
9. T. Maniatis, E. F. Fritsch, J. Sambrook, *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982).
10. C. Sund, J. Ylikoski, P. Hurskainen, M. Kwiatkowski, *Nucleos. Nucleot.* **7**, 655 (1988).
11. D. Pinkel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 9138 (1988).

12. We thank E. Johnsen for technical assistance and T. Hansson for molecular modeling. U. Pettersson offered critical comments on this manuscript. Supported by the Beijer, Procordia, and Borgström foundations; by NUTEK, the Technical and Medical Research Councils of Sweden; and by the Swedish Cancer Fund.

18 July 1994; accepted 1 September 1994

Localization of a Breast Cancer Susc ptibility G n , *BRCA2*, to Chromosom 13q12-13

Richard Wooster,* Susan L. Neuhausen,* Jonathan Mangion,* Yvette Quirk,* Deborah Ford,* Nadine Collins, Kim Nguyen, Sheila Seal, Thao Tran, Diane Averill, Patty Fields, Gill Marshall, Steven Narod, Gilbert M. Lenoir, Henry Lynch, Jean Feunteun, Peter Devilee, Cees J. Cornelisse, Fred H. Menko, Peter A. Daly, Wilma Ormiston, Ross McManus, Carole Pye, Cathryn M. Lewis, Lisa A. Cannon-Albright, Julian Peto, Bruce A. J. Ponder, Mark H. Skolnick, Douglas F. Easton,† David E. Goldgar, Michael R. Stratton

A small proportion of breast cancer, in particular those cases arising at a young ag , is due to the inheritance of dominant susceptibility genes conferring a high risk f the disease. A genomic linkage search was performed with 15 high-risk breast cancer families that were unlinked to the *BRCA1* locus on chromosome 17q21. This analysis localized a second breast cancer susceptibility locus, *BRCA2*, to a 6-centimorgan interval on chromosome 13q12-13. Preliminary evidence suggests that *BRCA2* confers a high risk of breast cancer but, unlike *BRCA1*, does not confer a substantially elevated risk of ovarian cancer.

In 1990, a breast cancer susceptibility gene, known as *BRCA1*, was localized to chromosome 17q (*1*). Subsequent studies demonstrated that *BRCA1* accounts for most families with multiple cases of both early-onset breast and ovarian cancer and about 45% of families with breast cancer only, but few if any families with both male and female breast cancer (*2*). Several other genes can confer susceptibility to breast cancer. Germline mutations in the p53 gene on chromosome 17p cause a wide range of neoplasms including early-onset breast cancer, sarcomas, brain tumors, leukemias, and adrenocortical cancer (*3*). Certain rare abnormalities of the androgen receptor appear to be associated with breast cancer in men (*4*), and epidemiological studies have suggested that heterozygotes for the ataxia telangiectasia gene, *AT*, on chromosome 11q are at elevated risk of breast cancer (*5*). However, mutations in *p53* and *AT* can only be responsible for a small minority of breast cancer families that are unlinked to *BRCA1* (*6*).

To localize other genes that predispose to breast cancer, we performed a genomic linkage search using 15 families that had multiple cases of early-onset breast cancer and that were not linked to *BRCA1*. These families were classified according to the number of cases of female breast cancer, male breast cancer, and ovarian cancer (Table 1). In addition to a negative lod score (logarithm of the likelihood ratio for linkage) with markers flanking *BRCA1*, all but one of the families used for this study had at least one breast cancer case diagnosed before age 50 that did not share a *BRCA1* haplotype with other breast cancer cases in the family. The exception, CRC 136, had an obligate sporadic case diagnosed at age 53. Families were genotyped with polymorphic microsatellite repeat markers (*7*, *8*). Typing of the markers *D13S260* and *D13S263* provided provisional evidence for the presence of a susceptibility gene on chromosome 13, which was subsequently confirmed by analysis of additional polymorphisms in the region.

R. Wooster, J. Mangion, Y. Quirk, N. Collins, S. Seal, M. R. Stratton, Section of Molecular Carcinogenesis, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.
S. L. Neuhausen, K. Nguyen, T. Tran, P. Fields, C. M. Lewis, M. H. Skolnick, D. E. Goldgar, Department of Medical Informatics, University of Utah, Salt Lake City, UT 84108, USA.
D. Ford, D. Averill, G. Marshall, J. Peto, D. F. Easton, Section of Epidemiology, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK.
S. Narod, Department of Medicine, Division of Medical Genetics and Division of Human Genetics, McGill University, Montreal, Canada H3G 1A4.
G. M. Lenoir, International Agency for Research on Cancer, 150 Cours Albert-Thomas, 69372 Lyon Cedex 08, France.
H. Lynch, Department of Preventive Medicine and Public Health, Creighton University School of Medicine, Omaha, NE 68178, USA.
J. Feunteun, Institute Gustav-Roussy, Villejuif, France.
P. Devilee and C. J. Cornelisse, Departments of Pathology and Human Genetics, University of Leiden, 2333 AL Leiden, Netherlands.
F. H. Menko, Department of Clinical Genetics, Free University of Amsterdam, 1007 MB Amsterdam, Netherlands.
P. A. Daly, W. Ormiston, R. McManus, Department of Medicine, Trinity College Medical School, St. James Hospital, Dublin 8, Ireland.
C. Pye and B. A. J. Ponder, CRC Human Cancer Genetics Group, Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK.
L. A. Cannon-Albright, Department of Internal Medicine, University of Utah, Salt Lake City, UT 84132, USA.

*These authors contributed equally to this study.
†To whom correspondence should be addressed.

nature genetics
volume 20 no. 3 november 1998

SNP attack on complex traits

Single nucleotide polymorphisms (SNPs) are major contributors to genetic variation, comprising some 80% of all known polymorphisms, and their density in the human genome is estimated to be on average 1 per 1,000 base pairs. Although SNPs are mostly biallelic—and consequently less informative than microsatellite markers—they are more frequent and mutationally more stable, making them suitable for association studies in which linkage disequilibrium (LD) between markers and an unknown variant is used to map disease-causing mutations. In addition, because SNPs have only two alleles, they can be genotyped by a simple plus/minus assay rather than a length measurement, making them more amenable to automation.

These are good reasons to develop SNPs as useful markers, but hardly sufficient to explain the momentum that the SNP movement has recently acquired, which stems from the hope that SNP-based approaches will lead to progress in the search for genetic variation associated with common diseases or sensitivity to drugs. At a recent meeting*, advances in SNP technology and SNP-based approaches to tackle complex traits as well as questions of human origin and prehistory were discussed. Frustrated with linkage analysis, which has had little success in identifying genes involved in determining complex traits, many geneticists have turned towards association studies which might be better suited to detecting genetic effects of low penetrance with higher resolution. For such studies, many more markers will be required—in addition to better statistical tools and high-throughput low-cost genotyping technology to analyse large marker sets in many samples.

Increasing amounts of sequence data available in public and private databases, (within which SNPs can be discovered *in silico*; Pui-Yan Kwok, Macdonald Morris), efforts underway to re-sequence DNA stretches from several individuals, and the use of 'SNP discovery' technology (such as denaturing high performance liquid chromotography; Peter Underhill), have led to the rapid accumulation of catalogued SNPs. So far, no SNP has been patented, but a number of applications are pending (Christian Stein), and it seems likely that many will end up in proprietary collections. Even with the best tools, understanding complex traits and human variation will be a challenge, to say the least; sharing resources will help. Two publicly available SNP databases as well as several SNP collections exist at present (see box)—and researchers are encouraged to submit any SNP that they discover.

The technological and economic goal is accurate, easy, cheap and fast large-scale SNP genotyping. Several methods are currently being developed, and it is unclear which one(s) will turn out to be the best. Examples based on minisequencing on DNA arrays (Ann-Christine Syvänen, Andres Metspalu), dynamic allele-specific

*First International Meeting on Single Nucleotide Polymorphism and Complex Genome Analysis. Skokloster, Sweden, 29 August–1 September, 1998, organized by Anthony Brookes, Ulf Landegren, Ann-Christine Syvänen, Anders Isaacson and Ulf Gyllensten, Uppsala University.

SNP databases

- **HGBASE (http://hgbase.interactiva.de)** collects intragenic SNPs and contains approximately 2,700 entries. It is searchable by sequence and, at the moment, the only database where information can be deposited and retrieved.
- **dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)**, a joint effort by the NHGRI and the NCBI, is now accepting submissions. Its curators are still working on making content available; the database will be searchable by STS accession number and fully integrated with GenBank.

SNP websites

- The **MIT SNP database (http://www-genome.wi.mit.edu/SNP/human/index.html)** contains over 3,000 SNPs (approximately two thirds of them mapped) and is searchable by genomic region or internal STS identifier.
- The **WashU SNP database (http://www.ibc.wustl.edu/SNP/)** contains several hundred SNPs which are currently being integrated into dbSNP.

hybridization (DASH, Anthony Brookes), microplate array diagonal gel electrophoresis (MADGE, Ian Day), pyrosequencing (Pål Nyrén), oligonucleotide-specific ligation (according to Ed Southern, the most sensitive assay) as well as the Whitehead/Affymetrix SNP chips (Jian-Bing Fan) and the TaqMan system (Ken Livak) were discussed. All of them require target amplification of each SNP by PCR. Even in the light of encouraging progress in multiplexing PCR (Michelle Cargill), a large number of individual reactions is required and the cost is considerable (James Weber). Ideally, one would like to determine the genotype directly from genomic DNA. Methods based on the generation of small signal molecules by invasive cleavage followed by mass spectrometry (Timothy Griffin) or immobilized padlock probes and rolling-circle amplification (Ulf Landegren) might eventually eliminate the need for PCR.

Apart from the challenges of generating SNP maps and efficient genotyping, how easy will it be to determine which SNPs are suitable for a particular question and how best to analyse the data? In the absence of understanding what makes complex traits complex, classical mendelian concepts (two alleles, normal *versus* abnormal) are usually imposed onto a more complicated reality. Joseph Terwilliger warned that only if the genes underlying complex diseases have one wild-type and one (or one major) susceptibility allele—that is, when allelic heterogeneity is low—is statistical analysis likely to detect association of the causative allele (or linked markers) with the disease phenotype. Intuitively, more markers should allow increased accuracy, but in statistical reality, this also means larger samples will be necessary or the risk of obtaining false positive results will increase. Skeptical about the use of SNPs in disease genetics, Terwilliger is nonetheless enthusiastic about their potential use in population genetics and genetic epidemiology. By way of contrast, Marta Blumenfeld and Nik Schork described a strategy by which they can overcome many of the statistical obstacles of SNP-based association studies. By sequencing DNA from a minimum of 100 individuals to establish SNP allele frequency, calculating LD strength in a region of interest prior to determining how many markers are needed, and analysing haplotypes (2–6 SNPs together) instead of individual markers, they have been able to identify new genes associated with complex traits—unfortunately the identities of the genes were not disclosed, and so proof of principle is yet to be provided.

Although the jury is still out on whether SNPs will provide easy answers to complex questions, they are increasingly popular with disease and population geneticists. While the former mainly concentrate on SNPs within or close to genes, the latter often prefer markers outside of genes (to avoid selection) and in areas of the genome devoid of recombination. Several approaches using SNPs on the Y chromosome (Chris Tyler-Smith, Francesc Calafell) and in a low-recombination interval on the X (Svante Pääbo) provide interesting leads on human history, as well as data about age, frequency and population distribution of SNPs. Of course, this is information directly relevant to disease geneticists, and underscores the need for more interaction between population and disease geneticists (Andrew Clark, Rosalind Harding). Knowledge about population evolution and history will reveal suitable populations for genetic studies and aid in study design and interpretation of results.

Time—or rather data—will tell whether SNPs live up to expectations. As Aravinda Chakravarti stated in his abstract, "Each genetic approach, considered either optimistic or pessimistic, has its underlying assumptions. Human geneticists have to begin to test these assumptions not by computer simulations and theoretical arguments but by empirical observations".

Publication 87 with RCE Submission July 18, 2003

09/623,068

only sequences in each specimen were examined by phylogenetic analysis.

11. The methods described by G. H. Learn *et al.* [*J. Virol.* 70, 5720 (1996)] were used to align DNA sequences (with the use of CLUSTALW plus manual adjustment), calculate genetic distances (with the use of DNADIST, using the maximum likelihood method), evaluate potential sample mixups, construct neighbor joining trees, and perform bootstrap analyses (1000 replicates). Sequence regions that could not be unambiguously aligned were removed from subsequent analyses. Each sequence was compared for phylogenetic relatedness to the entire set of published and available unpublished laboratory HIV database sequences. If after this analysis the viral sequences from a mother and an infant appeared as a monophyletic group on a phylogenetic tree, they were judged to be phylogenetically linked or to have a common ancestor not shared by sequences from any other individuals evaluated. Issues regarding the

12. Post-it® Fax Note 7671 | Date 7-14 | # of pages ▶ 6
13. To Marie | From Frances
14. Co./Dept. Burich | Co. Bell
15. Phone # | Phone #
16. Fax # | Fax #
17.
18.

19. T. Loy *et al.*, *J. Immunol.* 154, 3147 (1995).
20. A. Hoffenbach *et al.*, *ibid.* 142, 452 (1989).
21. G. Schochetman, S. Subbarao, M. L. Kalish, in *Viral Genome Methods*, K. W. Adolph, Ed. (CRC Press, Boca Raton, FL, 1996), pp. 25–41.

Kalish, J. W. Mos- *Retrovir.* 11, 1181

292 (1997).
g PCR assays: E. nan, L. M. Deme- arling, D. Shapiro, t and ACTG 076 cimons; D. Swof- PAUP", version K. K. Holmes for as supported by ndation (500153- 55525-ARI, 55532-ARI, 55526-ARI, 55531-ARI, and 55522-ARI), the U.S. Public Health Service (UO1-27658, AI32910, AI27757, and AI35539), and the Foster Foundation.

22 December 1997; accepted 26 March 1998

Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome

David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz,* Mark Chee, Eric S. Lander*

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome, and they provide powerful tools for a variety of medical genetic studies. In a large-scale survey for SNPs, 2.3 megabases of human genomic DNA was examined by a combination of gel-based sequencing and high-density variation-detection DNA chips. A total of 3241 candidate SNPs were identified. A genetic map was constructed showing the location of 2227 of these SNPs. Prototype genotyping chips were developed that allow simultaneous genotyping of 500 SNPs. The results provide a characterization of human diversity at the nucleotide level and demonstrate the feasibility of large-scale identification of human SNPs.

Although the Human Genome Project still has tremendous work ahead to produce the first complete reference sequence of the human chromosomes, attention is already focusing on the challenge of large-scale characterization of the sequence variation among individuals (*1*). This genetic diversity is of interest because it explains the basis of heritable variation in disease susceptibility, as well as harbors a record of human migrations.

The most common type of human genetic variation is the SNP, a position at which two alternative bases occur at appreciable frequency (>1%) in the human population. There has been growing recognition that large collections of mapped SNPs would provide a powerful tool for human genetic studies (*1*, *2*). SNPs can serve as genetic markers for identifying disease genes by linkage studies in families, linkage disequilibrium in isolated populations, association analysis of patients and controls, and loss-of-heterozygosity studies in tumors (*1*, *2*). Although individual SNPs are less informative than currently used genetic markers (*3*), they are more abundant and have greater potential for automation (*4*, *5*).

We performed an initial survey to identify SNPs by using conventional gel-based DNA sequencing to examine sequence-tagged sites (STSs) distributed across the human genome. STSs are short genomic sequences that can be amplified from DNA samples by means of a corresponding polymerase chain reaction (PCR) assay. From among 24,568 STSs used in the construction of a physical map of the human genome at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (*6*, *7*), an initial collection of 1139 STSs was chosen (*8*). These STSs contained a total of 279 kb of genomic sequence (*9*), with one-third from random genomic sequence and two-thirds from 3′-ends of expressed sequence tags (3′-ESTs) and primarily representing untranslated regions of genes. Each STS was amplified from four samples (*10*): three individual samples and a pool of 10 individuals (thereby permitting allele frequencies to be estimated among 20 chromosomes). The PCR products were subjected to single-pass DNA sequencing based on fluorescent-dye primers and gel electrophoresis; sequence traces were compared by a computer program followed by visual inspection (*11*). Candidate SNPs were declared when two alleles were seen among the three individuals, with both alleles present at a frequency greater than 30% in the pooled sample. The term "candidate SNP" is used because a subset of such apparent polymorphisms turn out to be sequencing artifacts, as discussed below.

The survey identified 279 candidate SNPs, distributed across 239 of the STSs. This corresponds to a rate of one SNP per 1001 base pairs (bp) screened and an observed nucleotide heterozygosity of $H = 3.96 \times 10^{-4}$ (Table 1). Expressed sequences (3′-ESTs) showed a lower polymorphism rate than random genomic sequence (with

D. G. Wang, C.-J. Siao, P. Young, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, E. Robinson, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
J.-B. Fan, A. Berno, R. Sapolsky, G. Ghandour, L. Hsie, T. Topaloglou, E. Hubbell, M. Mittmann, M. S. Morris, N. Shen, R. Lipshutz, M. Chee, Affymetrix, Incorporated, 3380 Central Expressway, Santa Clara, CA 95051, USA.
E. S. Lander, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*To whom correspondence should be addressed.

the difference falling just short of statistical significance at $P = 0.057$, one-sided), consistent with greater constraint within genic sequences. The ratio of transitions to transversions was 2:1. Although the dinucleotide CpG makes up only about 2% of the sequence surveyed, nearly 25% of the SNPs occurred at such sites with the substitution almost always being C↔T. Cytosine residues within CpG dinucleotides are the most mutable sites within the human genome, because most are methylated and can spontaneously deaminate to yield a thymidine residue (*12*). In addition to the single-base substitutions, 23 insertion-deletion polymorphisms were also found (with all but eight involving a single base), corresponding to a frequency of one per 12 kb surveyed.

Gel-based resequencing was satisfactory for the initial screen, but we sought a more streamlined approach for a larger scale SNP identification. One such approach involves hybridization to high-density DNA probe arrays (*13*). Such "DNA chips" can be produced with parallel light-directed chemistry to synthesize specified oligonucleotide probes covalently bound at defined locations on a glass surface or "chip" (*14*). A target DNA sequence of length L can be screened for a polymorphism by hybridizing a biotin-labeled sample to a variant detector array (VDA) of size 8*L* (Fig. 1). For each position on both strands, the array has four 25-nucleotide oligomer probes complementary to the sequence centered at the position. The four differ only in that the central (13th) position is substituted by each of the four nucleotides. Homozygotes (AA) for the expected sequence should hybridize more strongly to the perfectly complementary probe than to the three probes containing a central mismatch. The presence of an SNP would be expected to give rise to a different hybridization pattern, with homozygotes (BB) showing strong hybridization to an alternative base and heterozygotes (AB) showing strong hybridization to two probes. The VDA thus signals the presence of a sequence variation (by a change in the hybridization pattern) and, in many cases, indicates the nature of the change (by a gain of signal at a specific mismatch probe). VDAs have been used for mutation detection of small, well-studied DNA targets [such as 387 bp from the human immunodeficiency virus–1 genome, 3.5 kb from the breast cancer–associated *BRCA1* gene, and 16.6 kb from the human mitochondrion (*13*, *15*)] in large numbers of samples. In this setting, the normal hybridization pattern can be characterized with precision and single-base substitutions detected with high accuracy.

In this project, we used VDAs in a large-scale survey. A total of 16,725 STSs covering 2 Mb of human DNA were selected, with one-third from random genomic sequence and two-thirds from 3′-ESTs. The survey used 149 distinct chip designs, each containing 150,000 to 300,000 features. The STSs were examined in seven individuals, representing about 14 Mb of genomic sequence. For each chip, the corresponding STSs were amplified from an individual, pooled together, labeled with biotin, hybridized, and stained (*16*), and the resulting hybridization patterns were compared by a computer program followed by visual inspection (*17*). At each position, samples were classified as homozygous for the expected sequence, homozygous for an alternative sequence, or heterozygous.

A collection of 2748 candidate SNPs were identified, corresponding to a rate of one per 721 bp surveyed and an observed nucleotide heterozygosity of 4.58×10^{-4} (Table 1). The number of STSs containing SNPs was 2299. The SNPs had a mean heterozygosity of 33%, with the minor allele having a mean frequency of 25%. SNPs were found less often in 3′-ESTs than in random genomic sequence ($P < 0.023$, one-sided), consistent with greater constraint in genic regions.

The nucleotide heterozygosity rate was indistinguishable from the estimate obtained from gel-based sequencing ($P > 0.12$, two-sided test), as was the ratio of transitions to transversions and the proportion of SNPs occurring at CpG dinucleotides. SNPs were detected at a higher frequency in the chip-based survey because more samples were surveyed (seven versus three individuals). The observed increase of 38.8% (1/721 versus 1/1001) agreed closely

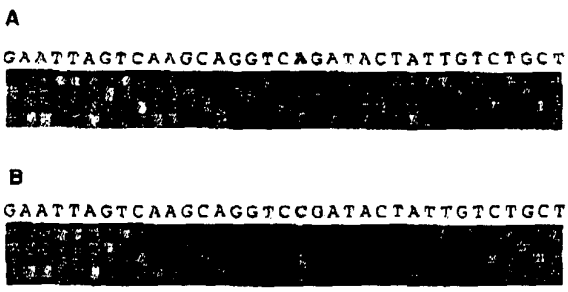


Fig. 1. SNP screening on chips. (**A**) Small portion of a VDA for an STS hybridized with the expected target sequence. Chip features in each column are complementary to successive overlapping 25-nucleotide oligomer subsequences, with the central base substituted by A, C, G, or T in the four rows. Variations from the expected sequence can usually be detected by examination of the most intense signal in each column. (**B**) The same VDA was hybridized with sequence containing an SNP (A→C) at position 19. The hybridization signal is now stronger at an alternative base at this position. It is also weaker at the surrounding positions (for example, positions 12 to 18 and 20 to 26), because probes at these positions are designed to be complementary to the A allele at the SNP and mismatch with the C allele.

Table 1. Results of SNP screening.

Variable	Gel-based sequencing			Chip-based detection		
	All STSs	STSs from 3′-EST sequences	STSs from random genomic sequence	All STSs	STSs from 3′-EST sequences	STSs from random genomic sequence
No. of STSs screened	1,139	705	434	16,725	12,649	4,076
Total bases screened	279,165	186,524	92,641	1,981,030	1,324,320	656,710
No. of candidate SNPs found	279	161	118	2,748	1,749	999
SNP frequency (*K*)	1/1001	1/1159	1/785	1/721	1/757	1/657
Heterozygosity (*H*) ($\times 10^{-4}$)	3.96 ± 0.38	3.42 ± 0.43	5.04 ± 0.67	4.58 ± 0.15	4.36 ± 0.18	5.02 ± 0.28
No. of STSs containing SNPs	239	137	102	2,299	1,515	784
% transitions among SNPs	67%	67%	67%	70%	70%	71%
% SNPs occurring within CpG	24%	23%	25%	24%	25%	22%
θ, based on *H*	3.96×10^{-4}			4.58×10^{-4}		
θ, based on *K*	4.33×10^{-4}			4.36×10^{-4}		

with expectation under classical population genetic theory (*18*). This result has implications for the choice of sample size for an SNP survey (*19*).

We estimated the error rates in the gel-based and chip-based surveys. The false-positive rate was estimated by carefully confirming candidate SNPs found in each survey by using thorough multipass sequencing (*20*): 12% of 220 candidate SNPs found in the chip-based survey and 16% of 120 candidate SNPs found by single-pass gel-based sequencing were false positives. The false-negative rate was estimated by considering a subset of STSs that had been included in both surveys: these STSs yielded 55 SNPs (all carefully confirmed to eliminate false positives), of which eight (15%) were missed by single-pass gel-based resequencing and seven (13%) were missed by the chip-based survey. Many of the errors were due to random factors, in that they were eliminated simply by repeating the original experiment. However, some were reproducible artifacts that could be eliminated only by changing the detection protocol (for example, by using dye terminators rather than dye primers in gel-based sequencing). The gel-based sequencing and chip-based analysis had similar rates of accuracy—with a false positive and false negative being found roughly every 5000 to 10,000 bases, or about 10% of the true SNP frequency. The accuracy largely reflects the particular implementation of the technologies in a high-throughput setting and could be increased at the expense of assay optimization.

Although the two surveys yielded comparable accuracy, the survey based on VDAs required considerably less laboratory work than gel-based resequencing. Both approaches required amplifying target loci. The gel-based approach then required a sequencing reaction and electrophoresis on each individual locus, whereas the chip-based approach allowed targets totaling 30 kb to be pooled into a single labeling reaction and hybridized (*21*).

The SNP collection from the two surveys was supplemented by two directed approaches based on public databases. First, we collected reports from the literature of common variants in gene coding regions. We were able to confirm 120 of 143 cases tested by virtue of detecting two alleles in our screening panel; the remainder may be true polymorphisms but simply monomorphic in the individuals tested. Second, the GenBank database contains multiple entries for some ESTs. Such entries were compared to identify single-nucleotide differences, which might reflect either common polymorphisms or sequencing errors in single-pass EST sequencing. We tested 200 such apparent differences and confirmed the presence of an SNP in 94 cases. These two directed approaches thus yielded an additional 214 SNPs.

The project has thus identified 3241 candidate SNPs to date. Confirmation (*22*) has so far been obtained for 1477 SNPs and is expected to yield ~2900 true SNPs. All information about the SNPs has been deposited on the Whitehead/MIT Center for Genome Research Web site (*www.genome.wi.mit.edu*) and will be updated with results of additional surveys and confirmation tests. The information is also being deposited in the GenBank database.

For SNPs to be useful in human genetic studies, they must be assembled into maps showing their chromosomal location. To create a third-generation map based on SNPs, we used whole-genome radiation-hybrid (RH) mapping (*6, 7, 23*), which infers the position of loci based on co-retention in a panel of human-on-hamster cell lines; it has become a primary method for constructing maps of the human genome (*6, 7*).

The current RH map of the human genome is anchored by a scaffold of 1036 genetic markers from an earlier genetic map consisting of simple sequence length polymorphisms (SSLPs) (*7*). SNPs can be integrated with respect to the earlier genetic map by determining their position on the RH map. We have localized 1880 STSs, containing 2227 of the 3241 candidate SNPs, on the RH map and thereby relative to the human genetic map (Fig. 2 and Table 2). SNPs are not evenly distributed among chromosomes or within chromosomes because most were derived from ESTs, which are known to have an uneven distribution (*6, 7*). SNP-containing STSs are present at a mean spacing of 2.0 centimorgans (cM) across the genome (*24*), and the map contains 58 intervals greater than 10 cM. The genetic distances on the map must be regarded as approximate because they are based on interpolation from distances in the RH map. It will be desirable to reestimate these distances on the basis of direct linkage analysis in the CEPH families, as high-throughput genotyping for the complete SNP collection becomes feasible.

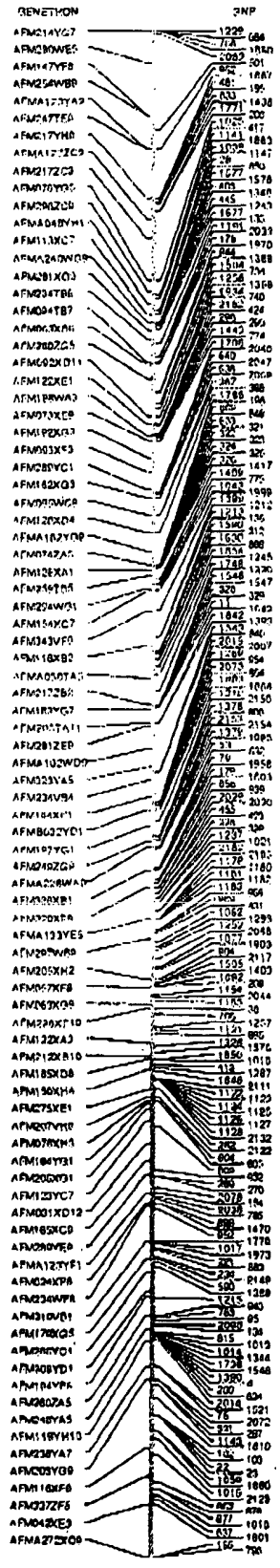


Fig. 2. A portion of the SNP genetic map (showing human chromosome 1). The full map is available on the Whitehead Institute Web site (*www.genome.wi.mit.edu*). Positions are based on genetic distances in centimorgans. Genetic positions of SNPs were inferred by localizing them relative to framework markers by RH mapping and then interpolating distances from centirays (on the RH map) to centimorgans (on the genetic map). Framework marker names are given in full. SNP names are named with the prefix WIAF (for example, WIAF-17), but the prefix is dropped and only the number is shown in the figure.

We next developed an efficient method for large-scale genotyping of SNPs based on extending the use of DNA chips from SNP discovery to SNP genotyping (5). We synthesized genotyping chips containing "genotyping arrays" for each SNP to be tested. Each genotyping array consists of two short VDAs corresponding to the two alternative alleles (Fig. 3). The presence of an allele should be reflected in strong hybridization to the corresponding resequencing array. PCR assays were designed for the region containing each SNP (25), with the goal of being robust and mutually compatible: the amplification targets were all small (typically, a few nucleotides around the polymorphic site), the primers all had similar calculated melting temperatures, and constant sequences were added to the 5'-ends of the forward and reverse primers to facilitate batch labeling of pooled PCR products. Each assay was tested to ensure that it amplified a single fragment from genomic DNA.

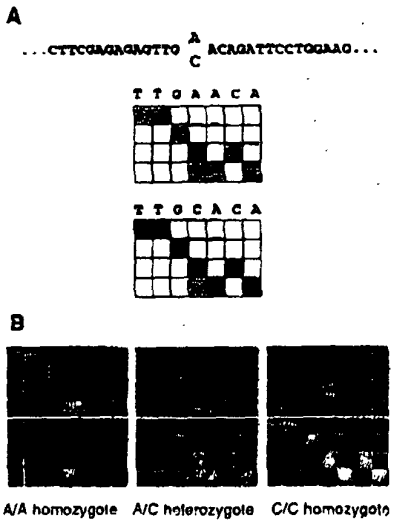


Fig. 3. Genotyping chips. (**A**) Schematic diagram of genotyping array for an SNP, consisting of two VDAs to study seven nucleotides centered around the SNP. The top and bottom arrays are designed to be complementary to the allelic sequences containing A and C, respectively. Probes perfectly matching the A and C alleles are shown in gray and black, respectively. A genotyping array for the complementary strand was also used but is not shown. (**B**) Hybridization signal for a genotyping array probed with samples from three individuals with respective genotypes AA, AC, and CC.

The most complex genotyping chip tested contained genotyping arrays for 558 candidate SNPs identified in the chip-based survey. Initially, the 558 loci were separately amplified, pooled, labeled, and hybridized to the chip. To determine whether each locus could be reliably read, we defined a formal detection test: loci passed if, for each of three individuals tested, the expected DNA sequence could be successfully read on both strands for one or both alleles. In all, 98% of the loci passed this detection test (with the remaining 2% failing as a result of weak hybridization or cross-hybridization).

We next sought to decrease substantially the sample preparation required to genotype large numbers of SNPs, as required to perform a genome scan. We developed a protocol based on multiplex PCR in which primer pairs from many different loci are combined in a single reaction (26). Although it is typically difficult to combine many PCR assays, the approach worked well for our SNP assays: 92% of the 558 loci passed the detection test when amplification was performed in 24 sets of ~23 loci; 90% passed when amplified in 12 sets of ~46 loci; 85% passed when amplified in 6 sets of ~92 loci; and 50% passed when amplified in a single set of 558 loci. The success appears to have resulted from a combination of factors, including the small size of the amplification targets, optimization of amplification conditions, and the presence of the constant sequence at the 5'-ends of the primers (27). It may be possible to salvage the unsuccessful assays by grouping them into additional multiplex sets or by redesigning the assays.

Multiplex amplification of sets of 46 loci was used in subsequent experiments because it decreased the number of reactions by a factor of 46 while allowing the vast majority (512/558) of loci to be assayed. The procedure was further tested in 39 individuals and was quite consistent: 96% of the 512 loci could be successfully read in 100% of individuals tested and the remainder in nearly all individuals.

We next developed a genotyping algorithm for each SNP. Loci were declared to pass a cluster test if the hybridization par-

Table 2. Chromosomal distribution of genetic markers.

Chromosome	No. of framework markers used from 5264-marker Genethon genetic map	Genetic distance in cM, on Genethon genetic map	No. of SNPs	No. of STSs	Avg. distance between STSs (cM)	No. of intervals >10 cM
1	88	293	236	201	1.5	3
2	91	277	177	149	1.9	2
3	78	233	160	133	1.8	1
4	54	213	98	87	2.4	4
5	65	198	86	72	2.8	4
6	71	201	156	118	1.7	4
7	57	184	119	94	2.0	1
8	45	166	135	108	1.5	3
9	40	167	106	88	1.9	3
10	53	182	85	78	2.3	1
11	58	156	105	92	1.7	1
12	43	169	108	91	1.9	3
13	23	118	57	45	2.6	4
14	31	129	83	64	2.0	3
15	29	110	76	67	1.6	0
16	36	131	70	64	2.0	0
17	23	129	94	87	1.5	2
18	29	124	52	43	2.9	4
19	22	110	58	53	2.1	2
20	34	97	58	49	2.0	2
21	18	60	29	26	2.3	2
22	12	58	31	28	2.1	1
X	36	191	46	43	4.4	8
Total	1036	3699	2227	1880	2.0	58

terns seen in a test set of 39 individuals fell into distinct clusters, corresponding to the possible genotypes (*28*). These clusters could then be used to assign genotypes for further samples (*29*).

The cluster test was applied to the ~500 candidate SNPs that worked well under multiplex amplification conditions: 75% passed the cluster test, and careful resequencing demonstrated that all such loci were true polymorphisms. The cluster test thus provides reliable confirmation of an SNP. The remaining 25% failed the cluster test, and resequencing revealed that half were false positives in the SNP screen and half were true polymorphisms (with the poor discrimination on the chip typically due to one allele hybridizing more weakly than the other). Thus, 88% of the candidate SNPs proved to be true polymorphisms, and 86% of true SNPs passed the cluster test.

To test the reproducibility and accuracy of the genotyping method, we genotyped a set of 91 loci (passing the cluster test) in three individuals by performing chip-based genotyping on six separate occasions over a 2-month period. The correct genotypes were independently determined by thorough gel-based resequencing. The genotyping-chip assay assigned a genotype in 98% of cases (1613/1636), and this assignment proved correct in 99.9% (1611/1613) of these cases. The loci were also genotyped in two complete CEPH families. The genotypes were not independently confirmed, but they were fully consistent with mendelian segregation.

For SNPs passing the cluster test, highly accurate genotypes could thus be obtained with the simple design used here. For the remaining SNPs (14%), similar accuracy can likely be obtained but may require optimization of the genotyping array design, depending on the locus [as shown in (*5*)].

The SNP surveys provide data about human genetic diversity. Two classical measures of diversity (*30*) are *H*, the average heterozygosity per nucleotide, and K, the proportion of sites harboring a variation. *H* does not depend on sample size, whereas *K* increases with the number of genomes surveyed. For a population at equilibrium, the neutral theory of evolution relates *H* and *K* to the classical population genetic parameter $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide. (θ can be thought of as twice the number of new mutations per generation arising in a population with size N_e.) Specifically, $H \approx \theta$ and $K \approx \theta[1^{-1} + 2^{-1} + 3^{-1} + \ldots + (n-1)^{-1}]$, provided that θ is small. From these equations, one can estimate θ based on *H* or *K*.

The human population is not at equilibrium, but rather underwent a rapid population expansion in the last 100,000 to 200,000 years. Such population explosions tend to suppress the effects of genetic drift and thus preserve the distribution of common alleles and the value of θ. Accordingly, the value of θ is relevant to the ancestral human population before its recent expansion.

The four estimates of θ derived from *H* and K for the two surveys are all roughly $\theta \approx 4 \times 10^{-4}$ (Table 1). Assuming a mutation frequency of $\mu \approx 10^{-8}$ to 10^{-9}, this would suggest an effective population size of $N_e \approx 10^4$ to 10^5, which seems reasonable for the ancestral population preceding the explosion in the last 100,000 years (*31*). Strictly speaking, these estimates apply only to the European population, from which all samples were drawn. However, a preliminary survey of a more diverse sample of 31 individuals representing all major racial groups yielded a value of θ that is only 30% larger (*26*), consistent with the idea that human variation occurs primarily within rather than between racial groups (*32*).

The resources reported here represent only a first step toward a dense SNP map of the human genome. The genetic map should already be useful for family-based linkage studies, given the average spacing (2 cM) and average heterozygosity (34%) of the markers. (The heterozygosity applies to the European-derived samples studied here, but a preliminary survey of ~180 of the SNPs shows that most are also polymorphic in other groups.) It still remains to develop a suitable genotyping system, such as a 2000-SNP genotyping chip.

Large-scale screening for human variation is clearly feasible. Someday it may become possible to screen entire human genomes. In the nearer term, a key goal will be to extend SNP discovery to the protein coding regions of all human genes (roughly 120 Mb of sequence, only about 40 times more than the current study) in order to catalog the common variants that may explain susceptibility to common genetic traits and diseases (*1*).

REFERENCES AND NOTES

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996); E. S. Lander, *ibid.* **274**, 536 (1996); F. S. Collins, M. S. Guyer, A. Chakravarti, *ibid.* **278**, 1580 (1997).
2. L. Kruglyak, *Nature Genet.* **17**, 21 (1997).
3. SNPs have only two alleles and are less informative than typical, multi-allelic simple sequence length polymorphisms (SSLPs). This disadvantage can be offset by using a greater density of SNPs: a genome scan with 1000 well-spaced SNPs, for example, will extract about the same linkage information as the current standard of 400 well-spaced SSLPs (*2*).
4. B. J. Conner *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 278 (1983); U. Landegren, R. Kaiser, J. Sanders, L. Hood, *Science* **241**, 1077 (1988); D. Y. Wu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2757 (1989); R. K. Saiki *et al.*, *ibid.*, p. 6230; A.-C. Syvanen *et al.*, *Genomics* **8**, 684 (1990); D. A. Nickerson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8923 (1990); K. J. Livak *et al.*, *Nature Genet.* **9**, 341 (1995); M. T. Roskey *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4724 (1996).
5. M. T. Cronin *et al.*, *Hum. Mutat.* **7**, 244 (1996).
6. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
7. G. D. Schuler *et al.*, *ibid.* **274**, 540 (1996).
8. STSs with the largest sizes were used in the gel-based screen, and the remaining STSs, having somewhat smaller sizes, were used in the subsequent chip-based screen.
9. The genomic sequence screened (279 kb) is the sum of the distances between the primer sites of the STSs successfully resequenced.
10. The individuals surveyed were chosen from Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees K104, K884, and K1331 from the Amish, Venezuelan, and Utah populations, respectively. The SNP survey by gel-based sequencing examined three unrelated individuals (K104-1, K884-2, K1331-1) and a pool of 10 individuals (K104-13, -14, -15, -16; K884-15, -16; K1331-12, -13, -14, -15). The SNP survey by chip-based analysis examined seven unrelated individuals (K104-1, -16; K884-2, -15, -16; K1331-12, -13).
11. STSs were amplified with their corresponding PCR primers as described (*6*), except that the forward primer was modified to include the M13 –21 primer site (5′-TGTAAAACGACGGCCAGT-3′) at its 5′-end. The resulting PCR products were subjected to dye-primer sequencing (*33*), with products detected on an ABI377 or ABI373 fluorescence sequence detector. Possible sequence variations were detected by the ABI Sequence Navigator software package, which suggests potential heterozygotes by identifying nucleotide positions at which a secondary peak exceeds a selected threshold (50%). Such apparent variations were then visually inspected to compare the patterns seen among the several individuals.
12. D. N. Cooper and M. Krawczak, *Hum. Genet.* **85**, 55 (1990).
13. M. Chée *et al.*, *Science* **274**, 610 (1996); M. J. Kozal *et al.*, *Nature Med.* **2**, 753 (1996).
14. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991); A.-C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994). The current generation of technology allows fabrication of 1.28 cm by 1.28 cm arrays of ~320,000 distinct oligonucleotides, each residing in a "feature" of ~20 μm by 25 μm and containing >10^7 copies of the probe.
15. J. G. Hacia *et al.*, *Nature Genet.* **14**, 441 (1996).
16. STSs were amplified with their corresponding PCR primers as described (*6*). PCR products intended for hybridization to the same chip (typically 100 to 200 STSs from a single individual) were pooled together for subsequent processing. About 1 to 2 μg of the pooled PCR product was purified with Qiaquick purification kit (Qiagen), fragmented with deoxyribonuclease (DNase) I (Promega) and labeled with biotin with terminal deoxynucleotidyl transferase (TdT, GibcoBRL Life Technology). The purification was performed according to the manufacturer's instructions. The fragmentation was performed in a 40-μl reaction with 0.2 unit of DNase I, 10 mM tris-acetate (pH 7.5), 10 mM magnesium acetate, and 50 mM potassium acetate at 37°C for 15 min, after which the reaction was stopped by heat inactivation at 96°C for 15 min. The terminal transferase reaction was performed by adding 15 units of TdT and 12.5 μM biotin-N6-ddATP (DuPont NEN) to the preceding reaction mixture, incubating it at 37°C for 1 hour, and then heat-inactivating it at 96°C for 15 min. The labeled samples were hybridized to the chip as follows. Samples were denatured at ~96°C for 5 to 6 min and cooled on ice for 2 to 5 min. Chips were first hybridized with 6× SSPET [0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA (pH 7.4), 0.005% Triton X-100] for ~5 min and then hybridized with the denatured sample in hybridization buffer [3M tetramethylammonium chloride, 10 mM tris-HCl (pH 7.8), 1 mM EDTA, 0.01% Triton X-100, herring sperm DNA (100 μg/ml), and 200 pM control oligomer] at 44°C for 15 hours on a rotisserie at 40 rpm. Chips were washed three times with 1× SSPET, 10 times

with 6× SSPET at 22°C, and stained at room temperature with staining solution [streptavidin R-phycoerythrin (2 μg/ml) (Molecular Probes) and acetylated bovine serum albumin (0.6 mg/ml) in 6× SSPET] for 8 min. After they were stained, the chips were washed 10 times with 6× SSPET at 22°C on a fluidics workstation (Affymetrix). Hybridization to the chip was detected by using a confocal chip scanner (HP/Affymetrix) with a resolution of 40 to 80 pixels per feature and a 560-nm filter.

17. Candidate SNPs were identified by using a combination of four algorithms followed by a visual inspection. At each position, the VDA contains one "expected" probe (corresponding to the sequence from which the chip was designed) and three "variant" probes (containing a substitution in the central position). The first algorithm (base-calling) looked for positions at which, in some individuals, a variant probe gave a stronger signal than the expected probe. The second algorithm (clustering) considered the signal vector s_{ij} from the eight probes at position i (four base substitutions on both strands) in individual j and looked for positions i at which the vectors s_{ij} fell into multiple clusters. The third algorithm (mutant fraction) was similar but focused only on the expected probe and a single variant probe at a time (rather than all three variant probes). The fourth algorithm (footprint detection) looked for the loss of signal that occurs at the expected probes in the neighborhood of an SNP (*13*, *15*). The algorithms have different sensitivities for detecting heterozygous and homozygous variations.
18. As discussed in the text below, the proportion K of polymorphic sites is expected to be proportional to $[1^{-1} + 2^{-1} + 3^{-1} + \ldots + (n-1)^{-1}]$, where n is the number of genomes sampled. The proportion of polymorphic sites is thus expected to increase by 39.3% when the number of genomes is increased from 6 (in the gel-based survey) to 14 (in the chip-based survey). This agrees well with the observed increase of 38.8%.
19. A relatively small sample size suffices to capture much of the common variation. The sample size of 14 has a 50% chance of detecting an allele with a frequency of 5%. Doubling the proportion of variant sites identified would require increasing the number of genomes surveyed from 14 to 335, on the basis of the formula for K. The larger sample size will tend to identify polymorphisms with lower heterozygosity.
20. STSs were resequenced on both strands with dye-primer and dye-terminator chemistry.
21. The chip-based approach has the further advantage that long STSs can be analyzed, whereas gel-based sequencing is limited to about 600 bp. It is thus possible to use fewer PCR products to analyze a region. The current study did not take advantage of this feature because we used short STSs already available from our previous work (*5*, *7*).
22. Confirmation was initially performed by multipass sequencing but is currently being done by using the clustering test on genotyping chips.
23. M. A. Walter *et al.*, *Nature Genet.* **7**, 22 (1994).
24. The lowest density occurs on chromosome X, which has the lowest density of STSs and which was screened in fewer total genomes in as much as the screening panel included three males.
25. For each SNP, PCR primers were chosen with the PRIMER software package (*6*) to closely flank the polymorphic base and to have a predicted melting temperature of 57°C. Forward and reverse primers were synthesized with the T7 and T3 promoter sites (5′-TAATACGACTCACTATAGGGAGA-3′ and 5′-AATTAACCCTCACTAAAGGGAGA-3′) at their respective 5′ ends. Each PCR primer pair was individually tested to determine if it produced a single clear fragment visible by agarose gel electrophoresis and ethidium-bromide staining, as described (*6*). PCR assays passing this test were further classified as being strong or weak according to the yield of the fragment produced. Primer pairs were grouped into multiplex sets, with the sets chosen to consist of either strong assays or weak assays.
26. Multiplex PCR was performed by using multiple PCR primer pairs in a single reaction. Specifically, multiplex PCR reactions were performed in a 50-μl volume containing 100 ng of human genomic DNA, 0.1 to 0.2 μM of each primer, 1 unit of AmpliTaq Gold (Perkin-Elmer), 1 mM deoxynucleotide triphosphates (dNTPs), 10 mM tris-HCl (pH 8.3), 50 mM KCl, 5 mM $MgCl_2$, and 0.001% gelatin. Thermocycling was performed on a Tetrad (MJ Research), with initial denaturation at 96°C for 10 min followed by 30 cycles of denaturation at 96°C for 30 s, primer annealing at 55°C for 2 min, and primer extension at 65°C for 2 min. After 30 cycles, a final extension reaction was carried out at 65°C for 5 min. Because the resulting PCR products were small, it was unnecessary to fragment them (as was done for the STSs in the SNP screen). The PCR products were then labeled with biotin in a standard PCR reaction, by using T7 and T3 primers with biotin labels at their 5′-ends. The reaction was performed with 1 μl of template DNA, 0.1 to 0.2 μM labeled primer, 1 unit of AmpliTaq Gold (Perkin-Elmer), 100 μM dNTPs, 10 mM tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM $MgCl_2$, and 0.001% gelatin. Thermocycling was performed with initial denaturation at 96°C for 10 min followed by 25 cycles of denaturation at 96°C for 30 s, primer annealing at 52°C for 1 min, and primer extension at 72°C for 1 min. After 25 cycles, a final extension reaction was carried out at 72°C for 5 min. The PCR products from the various multiplex reactions for an individual were then pooled together. One-tenth of the pooled sample was denatured and used for chip hybridization. Chips were hybridized, washed, stained and scanned, as above (*16*).
27. D. G. Wang, unpublished observations.
28. A classification procedure for assigning genotypes was derived for each locus on the basis of the hybridization results observed in a test population of 39 individuals. The proportions of the two alleles present in the i th sample, denoted $\pi_{A,i}$ and $\pi_{B,i}$ (with $\pi_{A,i} + \pi_{B,i} = 1$) were estimated essentially by comparing the observed hybridization signal to the expected signals for the two VDAs. The values $\pi_{A,i}$ for the 39 individuals lie in the interval [0,1] and should ideally cluster near 0, 0.5, and 1.0, but other patterns might occur because of differences in hybridization intensity between the two alleles. The values were optimally clustered (*33*) with the MODECLUS procedure of the SAS software package (SAS Institute). A maximum of three nonoverlapping clusters was permitted, defined by points with a minimum separation of 0.12. A locus failed the cluster test if all the samples fell into a single cluster, if the samples gave rise to two clusters but neither corresponded to the heterozygous genotype (AB), or if too many samples (more than 9 of 39) fell outside the three optimal clusters. A locus passing the cluster test gave rise to either three clusters (genotypes AA, AB, BB) or two clusters (genotypes AA, AB or BB, AB).
29. Subsequent samples were genotyped according to the cluster in which the hybridization pattern fell, with no genotype being called for samples falling outside these predefined clusters.
30. W.-H. Li, *Molecular Evolution* (Sinauer, Sunderland, MA, 1997); N. Takahata and Y. Satta, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4811 (1997); M. Nei and D. Graur, in *Evolutionary Biology*, M. K. Hecht, B. Wallace, G. T. Prance, Eds. (Plenum, New York, 1984), vol. 17, pp. 73–118.
31. F. J. Ayala, *Science* **270**, 1930 (1995).
32. R. C. Lewontin, in *Evolutionary Biology*, T. H. Dobzhansky, M. K. Hecht, W. C. Steere, Eds. (Appleton-Century-Crofts, New York, 1972), vol. 6, pp. 381–398.
33. M. M. DeAngelis, D. G. Wang, T. L. Hawkins, *Nucleic Acids Res.* **23**, 4742 (1995).
34. W. L. G. Koontz and K. Fukunaga, *IEEE Trans. Comp.* **C-21**, 171 (1972).
35. We thank D. Stern for construction of chip scanners used in the project, C. Chen-Cheng for computation work related to the polymorphisms among EST sequences in GenBank, T. Hawkins for sequencing of some STSs, and D. Lockhart for helpful comments on the manuscript. Supported in part by grants from Affymetrix, Millennium Pharmaceuticals and Bristol-Meyers-Squibb (to Whitehead Institute), from the National Human Genome Research Institute [to Whitehead Institute (HG00098) and Affymetrix (HG01323)] and from the National Institute of Standards and Technology [to Affymetrix (70NANBSH1031)].

5 February 1998; accepted 31 March 1998

RasGRP, a Ras Guanyl Nucleotide–Releasing Protein with Calcium- and Diacylglycerol-Binding Motifs

Julius O. Ebinu, Drell A. Bottorff, Edmond Y. W. Chan, Stacey L. Stang, Robert J. Dunn, James C. Stone*

RasGRP, a guanyl nucleotide–releasing protein for the small guanosine triphosphatase Ras, was characterized. Besides the catalytic domain, RasGRP has an atypical pair of "EF hands" that bind calcium and a diacylglycerol (DAG)-binding domain. RasGRP activated Ras and caused transformation in fibroblasts. A DAG analog caused sustained activation of Ras-Erk signaling and changes in cell morphology. Signaling was associated with partitioning of RasGRP protein into the membrane fraction. Sustained ligand-induced signaling and membrane partitioning were absent when the DAG-binding domain was deleted. RasGRP is expressed in the nervous system, where it may couple changes in DAG and possibly calcium concentrations to Ras activation.

The cellular properties of neurons are modulated by a number of extrinsic signals, including synaptic activity, neurotrophic factors, and hormones. These signaling systems alter the intracellular concentrations of second messengers such as calcium and cyclic nucleotides, and these small molecules can regulate the activities of protein kinases (*1*). As the mechanisms linking Ras signaling to nerve function are not completely understood, we developed a cDNA cloning approach to identify proteins that enhance Ras signaling in the brain. From rat brain mRNA, we derived cDNAs that